IMPERIAL COLLEGE LONDON

Department of Epidemiology and Biostatistics

SPACE-TIME EXPOSURE MODELLING OF TROPOSHERIC O_3 IN EUROPE

by

Fatima Yahya Al-Aidarous

A thesis submitted in fulfilment of requirements for

the degree of Doctor of Philosophy at Imperial College London

2013

To my two angels that changed my world:

Maryam and Alhassan

Abstract

Exposure models need to be developed which can be applied at the continental scale, while still reflecting local variations in exposure conditions. Land use regression (LUR) has been widely adopted to describe the spatial variations in air pollutants over the longer term but not for short-term time-variable exposures. This study, therefore, aimed to develop and validate a space-time O_3 model applicable to epidemiological studies investigating the health effects of short-term (e.g. daily) O_3 exposures at the small-area scale.

A geographical information system (GIS) was developed, incorporating data from 1211 O_3 monitoring sites across Western Europe and a range of predictors, stored as 100m grids, including land cover, roads, topography and meteorology. The spatial model consisted of a LUR model representing the long-term average for years 2001-2007. The monitoring sites were classified, using multivariate statistical techniques, into 13 site types based on a set of descriptive indicators, then 13 temporal models represented by time functions were produced – one for each site type. These were linked to the spatial model using probability of group membership as a weighting factor. Finally, local meteorological data were incorporated to produce the full space-time model to predict daily concentrations for point locations.

The spatial and temporal models were individually evaluated based on agreement with measurement data from a reserved subset of 20% of the monitoring sites. The performance of the spatial model was similar to other continental LUR models (R^2=0.67; RMSE=7.64 μg/m^3), while performance of the temporal models ranged from 0.3 to 0.5 (R^2). Including local meteorological data into the full spatial-temporal model improved correlation with the concentrations measured at 30 monitoring sites in the Netherlands (R^2= 0.42 without; R^2=0.53 with meteorology).

Modelling daily O_3 over large areas at a fine spatial scale is possible using this approach. Overall model performance was further improved as the temporal period was aggregated to weekly or monthly. The model was applied to mothers in two birth cohorts in the European Study of Cohorts for Air Pollution Effects (ESCAPE) to provide daily O_3 exposure estimates, which can be aggregated as needed to provide individualised exposures based on date of birth.

Copyright Declaration

Declaration of Originality:

I, Fatima Al-Aidarous, declare that the work presented in this thesis Space-time exposure modelling of tropospheric O_3 in Europe is entirely my own. Where any material or information has been derived from other sources, I confirm that this has been indicated and fully referenced in the thesis.

Signature of Author

Acknowledgments

There are many people I would to thank for their help and support throughout this thesis.

I wish to express my sincere gratitude to my thesis supervisor Prof. David Briggs who served as a constant source of help and encouragement to learn and absorb all the information during this period of time. This PhD would have been impossible without his countless ideas and suggestions and unlimited support.

I wish to express my special gratitude to co-supervisor Dr. Danielle Vienneau for her encouragement and guidance over the years.

Special thanks go to group of people at Imperial College, Dr. Marta Blangiardo for answering thousands of questions in statistics and Juan G Gonzalez, as well as Margaret Douglass for her programming expertise. My thanks also go to Chloe Morris for her help and friendship, and everyone else who helped me during these four years.

A very special word of thanks is due to my parents for their support and encouragement, for my children for their tolerance when I could not spend the vacation times with them, and special thanks to my husband for his invaluable help and support throughout this study.

Finally, I am indebted last but not least to the Scholarship Coordination Office, for the financial support granted by The Ministry of Presidential Affairs, UAE under the patronage H.H Sheikh Zayed Al-Nahyan and continuing under H.H Sheikh Khalifa Bin Zayed Al-Nahyan, President of the United Arab Emirates.

Table of Contents

List of Tables

List of Figures

Abbreviation

AIRBASE	air quality information system database	
ANOVA	Analysis of variance	
APMoSPHERE	Air Pollution modelling for support to policy on health and environmental risk in Europe	
ARIMA	autoregressive integrated moving average	
AT	alert thresholds	
CAFF	Clean Air for Europe	
CORINE	coordination of information on the environment	
CLC2000	CORINE land cover data	
CMAQ	The Community Multiscale Air Quality	
SAPALDIA	Swiss study on Air Pollution and Lung Disease in adults	
DA	discriminant analysis	
EC	European Commission	
ECMWF	European Centre for Medium-Range Weather Forecast	
EEA	European Environment Agency	
ESCAPE	European Study of Cohorts for Air Pollution Effects	
EU	European Union	
EUROAIRTNET	European Air Quality Monitoring Network	
FB	Fractional Bias	
FEV	forced expiratory volume	
FVC	Forced vital capacity	
GIS	geographical information system	
USA	United states of America	
HCA	hierarchical cluster analysis	
HIA	Health impact assessment	
IDW	Inverse distance weighting	
IT	the information	
LAEA	Lambert Azimuthal Equal Area projection	
LOOCV	A leave-one-out cross-validation	
LUR	Land use regression	
MLOR	multinomial logistic modelling	
NMSE	normalised mean square error	
Pollutant:	CO	carbon monoxide
	HC	hydrocarbons
	NO	nitric oxide
	NO2	nitrogen dioxide
	Nox	nitrogen oxide
	O_3	ozone
	PM_{10}	particulate matter
	SO_2	sulphur dioxide
	VOC	volatile organic carbons
	CH_4	methane

PCA	principal component analysis
TV	target value
UNECE	United Nations Economic Commission for Europe
VAC	Variance component analysis
WHO	world health organization

PART 1: FOUNDATION

1 Introduction

1.1 Rationale

Ozone (O_3) pollution in the lower atmosphere (troposphere) has been an issue of considerable policy concern for many years. Early concerns mainly focused on the potential for damage to ecosystems, but by the early 20th century risks to human health were also recognized, and in 1992 the European Union's O_3 Directive established both guidelines and short-term warning thresholds. In 2002, further action was taken to control the emission of O_3 precursors, and the Clean Air for Europe (CAFE) programme has led to implementation of a broader and more encompassing policy on air quality in Europe under the Sixth Environmental Action Plan[1].

Climate change is also likely to increase O_3 levels in the atmosphere over the next century (Meleux et al., 2007) because there is a strong relationship between temperature and ambient O_3 concentrations. A statistical analysis in south and central Europe, for example, showed that between 1993-6 and 2000-4, the number of days in which O_3 concentrations exceeded the threshold of 120μg/m^3 increased by 8 days/year, as a consequence of the general temperature trend (EEA, 2008). Changes were most evident in urban areas: the same report noted that, while there was no change in O_3 concentrations in the rural areas from 1990-2007, in urban and trafficked areas there was a continuing upward tendency. Approximately 83% of the monitoring stations in European countries reported one or more exceedance of the threshold of 120μg/m^3 in summer of 2007.

The increase in O_3 concentrations is of special concern due to its adverse health effects as determined through both toxicological studies (Hazucha and Lefohn, 2007, Cotgreave, 1996, Mustafa, 1990) and epidemiological studies (Le et al., 2012, Karakatsani et al., 2010, Hathout et al., 2006, Park et al., 2005, Salam et al., 2005, Brook et al., 2002a, McDonnell et al., 1999).

Many of these epidemiological studies have used relatively simple measures of exposure. Most time-series studies have assessed exposures on the basis of the measured concentrations at the nearest monitoring site. Given the sparse distribution of the monitoring networks, this has meant that large portions of the entire study population may be assigned to a single site. Inevitably this causes substantial exposure misclassification, by ignoring intra-urban variations in concentrations (Wilhelm et al., 2009).

[1] http://ec.europa.eu/environment/newprg/intro.htm last accessed:20th April2012.

The limitations of the often sparse monitored data highlight the need to develop better techniques for modelling O_3 concentrations for exposure assessment. Because of the episodic nature of O_3 pollution, ideally these need to take account of temporal, as well as spatial, variations in concentrations. Because of the growing need to estimate exposures over large study populations (either to support large epidemiological studies or for the purpose of risk assessment and management), models also need to be applicable at the continental scale – whilst still reflecting local variations in exposures. Few studies have yet attempted to map O_3 concentrations at this scale and spatial resolution (Beelen et al., 2009). This suggests that, using appropriate techniques, it should be possible to derive high resolution maps of O_3 concentrations to facilitate and increase the effectiveness of exposure assessment.

As this brief introduction shows, and as stated by (Hoek et al., 2008), there is a need to develop air pollution exposures for health studies taking account of temporal variations on a fine temporal scale. The combination of time with spatial dimensions inevitably greatly increases the statistical challenges in the modelling. On the other hand, exposure estimation is only likely to be reliable if it allows for both temporal and spatial variations in concentrations. By the same token, improving exposure assessment in this way should improve the accuracy and sensitivity of studies designed to evaluate health risks from exposures to O_3. This will be the focus, and challenge, of this research as outlined in the aims below.

1.2 Aims and objectives

1.2.1 Aims

The main aim of this study is to develop a powerful, GIS-based methodology for modelling spatial and temporal O_3 concentrations using a combination of land use regression model (LUR) and Fourier analysis techniques over a large study area, as a basis for estimating the health impacts in long (i.e. weeks, months and years) and short-terms(i.e. days) studies. To optimise its usability, this methodology will make use of readily available data (both on ozone concentrations and the factors that determine them) and will be applicable at a range of both spatial and temporal scales.

1.2.2 Objectives

The specific objectives of this thesis are as follows:

1) Explore the spatial and temporal variations in monitored O_3 concentrations across western European countries, in order to assess the importance of different components of variation, including year, season, day of week, hour of day, region/country and 'site classification', as a basis for devising an appropriate modelling strategy.

2) Categorise the O_3 monitoring sites on the basis of the temporal characteristics of their O_3 concentrations, using a set of indicators representing the main elements of variation deduced from objective 1.

3) Develop an environmentally based zonation of site type that discriminates between these different site types, as a basis for extrapolating the measurements across the study area.

4) Develop and validate a spatial model of long term mean O_3 concentrations using GIS (i.e. LUR) techniques.

5) Develop and validate a temporal exposure model, at hourly and daily level, by fitting trigonometric functions, to the measured concentration data in each site type.

6) Combine the LUR and time function models to provide a space-time model of O_3 concentrations, and validate this against monitored concentrations at a reserved set of monitoring sites.

7) Explore the potential to enhance this space-time model by incorporating additional, daily information on meteorology.

1.3 Structure of thesis

The thesis is organised as follows:

- Chapter 2 presents a literature review related to: O_3 chemistry and production; regulatory guidance on O_3 concentrations; O_3 trends in Europe and the 'state of the art' of O_3 monitoring and modelling; health effects of O_3 exposure.

- Chapter 3 outlines the overall logic of, and steps in, modelling used in the thesis, and details the geographic information system (GIS) developed for this study, including description of the data sets used to define the variables used for modelling.

- Chapter 4 focuses on the classification of the O_3 monitoring sites, using hierarchical cluster analysis (HCA) of a set of indicators created to reflect temporal variations in concentrations, followed by determining the relationship between these site types and selected environmental variables, using multinomial logistic modelling (MLOR) to estimate the probability of membership of all site types at each unmonitored location.

- Chapter 5 describes the development of the 100m resolution LUR model for Europe, as a basis for mapping long term mean O_3 concentrations, and validation of the LUR using external data.

- Chapter 6 describes the methodology for creating a time function model (TM), for each site type, using trigonometric functions in Fourier analysis and presents and discusses the results.

- Chapter 7 explains the construction of the space-time model, combining the LUR (spatial model) with the TMs. It also presents two case studies in which the model is applied to estimate exposures for members of existing cohorts at different spatial scales (local and country level).

- Chapter 8 summarises the key findings of the research and discusses the potential sources of uncertainty in developing the space-time O_3 model and the lessons learned from this work. It also includes discussion about the opportunities presented by this type of model, and implications for exposure assessment.

2 State of the Art

2.1 Ground level O_3 and health impacts

Based on the latest European Environment Agency technical report (EEA, 2011), O_3 and particulate matter (PM) are Europe's most problematic pollutants in terms of health impacts. Partly for this reason, but partly also because of the potential impacts on vegetation, considerable attention has been given to the problem of O_3 pollution in recent years. This has helped to improve understanding of the mechanisms of O_3 production and dissipation, and the need for policies to control O_3 concentrations in the troposphere. This section outlines the formation of ambient O_3 in the troposphere, current air quality guidelines for its control, and the potential health impacts from exposure to O_3.

2.1.1 O_3 production in the troposphere

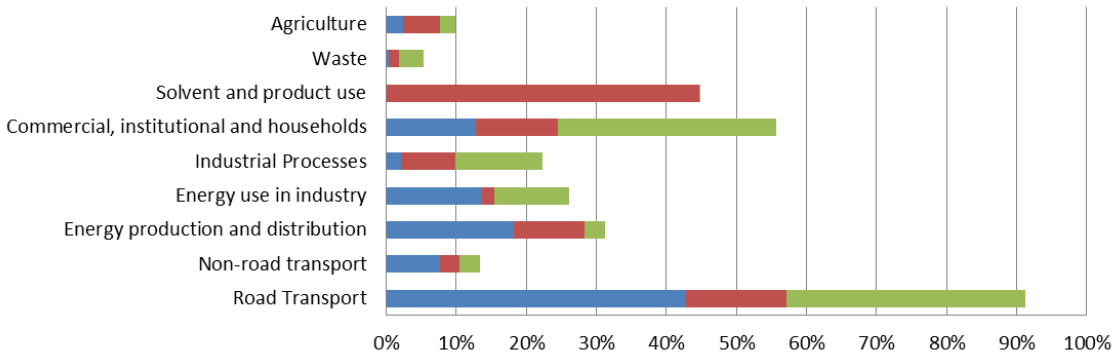
Research on O_3 chemistry and formation has expanded since the 1950s, partly in response to a major pollution problem in Los Angeles in the late of 1940s (Finlayson-Pitts and Pitts Jr, 2000a). This problem, which led to extensive incidents of human eye watering and plant death, was noticed during periods of sunshine and high outdoor temperature. Laboratory experiments, involving the exposure of plants to a range of hydrocarbons (HC) and nitrogen oxides (NOx) in the presence of sunlight replicated these symptoms and pointed towards some interaction between air pollutants and sunlight as the cause. Subsequently, O_3 was pinpointed as a major agent in this process (Mills, 1957).

O_3 is a secondary pollutant formed from a series of photochemical reactions between nitrogen oxides and volatile organic carbons (VOC) in the presence of sunlight. NOx, carbon monoxide (CO), non-methane volatile organic carbons (NMVOCs) and methane (CH_4) contribute to different extents to O_3 formation, but the most important O_3 precursors are NOx and NMVOCs. As Finlayson-Pitts and Pitts Jr (2000b) reported, CH_4 does not contribute significantly to O_3 formation due to it slow oxidization in the troposphere; therefore the term VOCs used here refers to volatile organic compounds including only NMVOC present in the gaseous phase in the troposphere. VOCs result from human activities such as road transport, solvent use, industrial processes, energy production and distribution, waste disposal, and agriculture. Natural processes are also responsible for a

substantial amount of VOC including emission from plants, trees, animals, and bacterial processes in soils (Derwent, 1995). VOC from natural sources - e.g. isoprene from deciduous trees and monoterpenes from conifers - are more reactive than the VOC emitted from human activities (Derwent, 1995, Sillman, 1999). Figure 2.1 shows that the largest proportion of VOC emissions in Europe in year 2009 are from solvent and product use (45%), followed by road transport (15%).

NOx includes nitrogen dioxide (NO_2) and nitric oxide (NO), and both gases are mainly emitted from the human activities listed in the table of Figure 2.1 along with natural processes such as lightning, forest fires, and bacterial processes in soil. In general, road transport is the most important source for NOx in Europe, accounting for 43% of the emissions, followed by energy production and distribution (18%). CO is produced from incomplete fuel combustion and also from the natural biological processes in soil and plants. Most CO in Europe is emitted from road transport (34%) and commercial, institutional, and households (31%).

In general, human activities such as road transport and highly populated built up areas (with industrial, commercial and domestic sources) are expected to contribute to high levels of O_3 precursors.



	Road Transport	Non-road transport	Energy production and distribution	Energy use in industry	Industrial Processes	Commercial, institutional and households	Solvent and product use	Waste	Agriculture
NOx	42.7%	7.7%	18.2%	13.7%	2.1%	12.8%	0.0%	0.4%	2.4%
VOCs	14.5%	2.7%	10.1%	1.7%	7.8%	11.7%	44.7%	1.5%	5.3%
CO	34.1%	2.9%	3.0%	10.7%	12.3%	31.2%	0.0%	3.4%	2.3%

Figure 2.1 Percent O_3-precursor emissions by source in EEA member countries (EU-152) based on 2009 data[3]

For the definition of each source see Appendix A, Section II

[2] Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden and the United Kingdom

[3] http://www.eea.europa.eu/data-and-maps/indicators/emissions-of-O3-precursors-version-2/assessment-1

As summarised in Sillman (1999,2003) and Finlayson-Pitts and Pitts Jr (2000a), the interactions of these precursors in the formation of O_3 involve the following reactions:

$VOC + OH \xrightarrow{[O2]} RO_2 + H_2O$	**Equation 2-1**
$CO + OH \xrightarrow{[O2]} HO_2 + CO_2$	**Equation 2-2**
$NO + HO_2 \rightarrow NO_2 + OH$	**Equation 2-3**
$NO_2 + h\nu \rightarrow NO + O$	**Equation 2-4**
$O + O_2 + M \rightarrow O_3 + M$	**Equation 2-5**
$NO + O_3 \rightarrow NO_2 + O_2$	**Equation 2-6**

These processes start with the reaction of VOCs or CO with the OH radical[4] (equations 2-1 and 2-2). The resulting RO_2 or HO_2 radicals[5] then convert NO to NO_2, generating further OH radicals (equation 2-3). Through photolysis (*hv*), NO_2 produces atomic oxygen (O) which further combines with O_2, in the presence of other molecules (M), to form O_3 (equations 2-4 and 2-5).

During the night, in the winter season, and in areas of high NO emission (e.g. transportation corridors and power plants), NOx-titration occurs (removal of O_3). In these situations equation 2-6 is dominant over equations 2-4 and 2-5. In the daytime, however, reaction 2-6 is balanced by the former reactions (equations 2-4 and 2-5).

The above-mentioned reactions show that formation of O_3 is controlled by the rate of the initial reaction of VOC and OH radicals (equation 2-1) and additionally the rate of NO and NO_2 emissions (equations 2-4 to 2-6), as well as the presence of sunlight. Therefore, the relationship between O_3, NOx and VOC is determined by complex photochemistry reactions. The isopleth in Figure 2.2 shows the O_3 concentration (in ppb) as a function of NOx and VOC emission rates. The dashed line represents the transition from VOC-sensitive to NOx-sensitive conditions and generally follows a line of constant VOC:NOx.

[4] Radical is formed mainly from photolysis of O_3 followed by photolysis of nitrous acid and hydrogen peroxide.
[5] Peroxide radicals: Number of organic chain attached to O_2 (replacing H the original chain) as when propane (C_3H_8) react with OH to produce $C_2H_7O_2$.

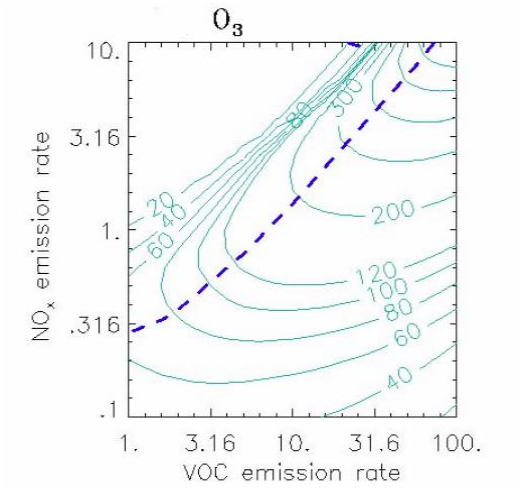


Figure 2.2 Isopleth plot of O_3 concentrations in ppb from (Sillman, 2003)

There is no simple "rule of thumb" for distinguishing between these two conditions in NOx-VOC chemistry. In general, however, NOx-sensitive conditions occur when there is relatively low NOx and high VOC which usually occurs in remote areas. On the other hand, VOC-sensitive conditions arise when there is a relatively high NOx and low VOC concentration, as typically occurs near to power plants and transport corridors. As Figure 2.2 shows, under VOC-sensitive conditions, O_3 increases in response to an increase in VOC, while in NOx-sensitive conditions, O_3 increases as NOx emissions rise.

NOx-VOC chemistry thus varies from location to location and time to time. Rural areas, for example, tend to be more NOx-sensitive than urban ones, but in the autumn may become VOC-sensitive. Large urban areas appear to be predominantly VOC-sensitive but NOx-sensitive chemistry may develop at locations downwind of major sources.

In summary, O_3 precursor concentration is affected both by the source characteristics and weather conditions. The formation of O_3 differs from place to place, and in any particular location may vary over time depending on the NOx-VOC-chemistry and conditions.

2.1.2 Regulatory guidance for the control of O_3 in Europe

As mentioned in Section 2.1.1, the effects of O_3 pollution were first recognised in the 1950s in Los Angeles. By the 1970s O_3 was recognised in Europe as a pollutant of concern (TRS, 2008), and in the late 20th century, regulatory guidelines were established to control emissions of O_3 precursors. A

number of policies have since been introduced in Europe, aimed at reducing the release of O_3 precursors, and maintaining safe levels of ambient concentrations for human health and vegetation.

O_3 precursor emissions are regulated by a number of directives and standards, for instance the VOC Solvent Directive (e.g. 1999/13/EC), the European standards for road vehicles (e.g. Regulation (EC) No 715/2007), and a number of international protocols (e.g. Gothenburg Protocol 1999) under the United Nations Economic Commission for Europe (UNECE) Convention on Long-range Trans-boundary Air Pollution (LRTAP convention), designed to cut emissions of O_3 precursors[6].

These regulatory guidelines are defined to control the effects of O_3 either on vegetation or human health, or both. The most important in terms of health is the European Union Ambient Air Quality Directive 2008/50/EC (EC, 2008) which entered into force on 11 June 2008, replacing the 2002 (2002/3/EC) O_3 Directive. The current Directive sets out four O_3 indicators for human protection, listed in Table 2.1. These include a long term objective and target value as well as the information and alert thresholds.

Table 2.1 O_3 guideline limits as set out in EU Air quality Directive 2008/50/EC

Objective	Level(μg/m³)	Averaging time
Long-term objective	120	Maximum daily 8-hour mean within calendar year
Target value	120, not to be exceeded more than 25 days per calendar year averaged over three years	Maximum daily 8-hour mean
Information threshold	180	One hour mean
Alert threshold	240	One hour mean

For health protection, a value of 120 μg/m^3 for the maximum daily 8-hour mean was set as the target value and long term objective. The target value aims to avoid, prevent or reduce harmful effects on human health, and is to be attained where possible over a given period (i.e. daily 8-hour mean). In the O_3 Directive, the target value of 120 μg/m^3 should not be exceeded more than 25 days per year, averaged over a three year period. The long term objective refers to the level to be attained in the long term, again with the aim of providing effective protection of human health. The long term objective is defined as the O_3 concentration, according to existing scientific knowledge, below which any direct adverse health effects are not expected. For any given calendar year, mean O_3 should not exceed 120 μg/m^3.

More recently, the WHO has updated the air quality guideline for O_3 based on a review of further evidence. It should be noted, however that, unlike the EU Directive, the WHO guideline is not

[6] http://www.unece.org

mandatory in Europe. An air quality guideline of 100µg/m³ for a daily maximum 8-hour mean is defined as the limit needed to protect the general population. This guideline limit is based on new chamber and field studies, which indicate that adverse health effects are liable to occur even if the EU target value limit of 120µg/m³ was met (WHO, 2006).

Additional thresholds have been set in the EU Directive to warn the public of high O_3 concentrations. The information threshold is defined as the concentration of O_3 which constitutes a risk to human health, especially the sensitive population (i.e. children, pregnant women, the elderly and people with respiratory diseases). Any exceedance of the information threshold (hourly O_3 concentration above 180 µg/m^3) should be reported to the EC by the Member State in which it occurred. The alert threshold is defined as an hourly O_3 concentration of 240 µg/m^3, and is intended to represent the concentration of O_3 above which constitutes a risk to human health for the general population. In this situation national authorities must warn the public and give advice[7].

2.1.3 O_3 trends in Europe

The extent to which these standards and guidelines are currently met varies considerably, both between countries and over time. Between 1997 and 2009, the percentage of the urban population in Europe exposed to O_3 concentrations above the 120 µg/m³ level rose from 13% to 61%, with most occurrences in Southern Europe. Also, 95% of the total urban population in Europe was exposed to O_3 concentrations above 100 µg/m³ between 2006-2008 (EEA, 2011). Extreme concentrations also seem to have become more extensive. The highest one hour mean ambient O_3 concentrations recorded in 2001, for example, were 360 µg/m³ in Spain and 387µg/m^3 in France (EEA, 2001); in 2008 the maxima reported were 302-399µg/m³ in Italy and 240-300µg/m^3 in Belgium, Greece, Italy, Spain and Switzerland (EEA, 2009). Exceedances of the threshold concentration of 120 µg/m³ over a two day period were reported at 28% of Europe sites in 2008 compared to 10% in 2001. Moreover, exceedance of the Long-term objective value of 120 µg/m³ has been fluctuating somewhat during the period 1997 to 2011 as demonstrated in Table 2.2. These data also clearly show that O_3 trends are affected by meteorological factors, with the greatest exceedances in both of the recent heat episode years, 2003 and 2006.

[7] http://www.eea.europa.eu/maps/O3/resources/faq/what-is-the-difference-between-information-threshold-and-alert-threshold

Table 2.2 Exceedances of the EU long term O_3 objective of 120 µg/m³ for the protection of human health during summer in Europe (1997-2011), from (EEA, 2012)

Summer season	No. of stations	Percentage of stations reporting exceedances(a)	Percentage of stations reporting exceedances(b)	Maximum observed 8-hour concentration (µg/m³)
1997	756	92	28	243
1998	811	91	31	263
1999	1138	93	30	537
2000	1206	92	29	242
2001	1368	92	39	269
2002	1421	89	30	310
2003	1510	95	68	296
2004	1545	91	27	256
2005	1667	92	34	291
2006	1764	95	53	399
2007	1795	87	31	277
2008	1905	90	22	399
2009	1921	89	23	244
2010	2193	85	27	262
2011	2186	84	24	259

(a) The number and percentage of stations at which at least one exceedance was observed
(b) Above 25 days

In principle, the long-term trend in Europe might be expected to be downwards, due to increasing controls on the release of O_3 precursor gases, under EU policies. In practice, the picture is much less clear. Wilson et al. (2012), for example, analysed O_3 trends in Europe from 1996 to 2005 based on observations of 158 rural sites and reported an increase in annual mean O_3 concentration of 0.16 ppb each year. Similar findings were reported in the latest EEA report (2011) which says that, despite the reduction in the emissions of O_3 precursors between 1999 and 2009, no corresponding drop in annual mean O_3 concentrations (other than a reduction in peak concentration) could be detected. This might be due to a number of factors, including uncertainties in the emission data and the complex relationship between precursor emissions and O_3 concentrations (NOx-VOC chemistry). An earlier report, however, indicates that, during the summer of 2007 and 2008, exceedances of the O_3 threshold in Europe were low compared to 1997 due to a reduction in peak values (EEA, 2008). Median concentrations, however, showed an increasing trend, more prominently at traffic and urban monitoring sites; and at the most polluted sites, winter concentrations were tending to increase, apparently due to a reduction in titration by NOx emissions (Jonson et al., 2006).

Comparisons between different years are nevertheless difficult, due to changes in the number and location of monitoring stations, and in the reported indicators. Interpretation of the trends is also made difficult by the poor spatial representativeness of the monitoring networks, and the low

spatial resolution of the data (Derwent et al., 2005). As it now stands, therefore, O_3 trends in Europe are not entirely understood. As these statistics also indicate, ambient O_3 concentrations vary substantially over time and these fluctuations occur at different amplitudes – e.g. annual, seasonal, and daily – in response to changing rates of O_3 formation and destruction.

2.1.4 Health impacts of O_3

Exposure to ambient O_3 in the lower atmosphere may lead to harmful consequence on human health that may occur either in the long-term or short-term. O_3 may be related to different types of morbidity, such as respiratory, cardiovascular and adverse birth outcomes, and in severe cases it may lead to mortality. These effects will be detailed below, in sections focusing on short term and long term health effects.

Controlled laboratory studies and studies exploring the effects of exposure to mixtures of air pollutants have suggested that the effects of exposure to ambient O_3 are independent from those of other air pollutants such as PM (UNECE, 2008). Over recent years, a number of epidemiological studies have investigated associations between atmospheric O_3 and human health, using a variety of study designs. Compared to studies of other regulated pollutants, such as PM and NOx, however, research is still relatively sparse.

2.1.5 Short term health effects

Most of the epidemiological studies of O_3 to date have focused on short-term health effects, and have been either cross-sectional or time series in design. According to the Task Force on Health (UNECE, 2008), the majority of recent epidemiological studies have reported positive and significant associations between short-term exposure to different O_3 concentrations and increased morbidity and mortality from respiratory diseases. Inhaling O_3, in the short term, can cause a variety of health problems, including lung damage, aggravated asthma, and increased susceptibility to respiratory tract illnesses such as pneumonia and bronchitis. The most consistent associations have been seen with impaired pulmonary function (WHO, 2008); for this outcome, increasing exposures to ozone were found to be correlated with increased medication usage (UNECE, 2008).

A 2005 report by the WHO to update air quality guidelines evaluated all available evidence on the health impacts of ozone exposure, and this was further updated in 2008. Most studies considered in

the review identified significant positive associations between short term increases in ambient ozone and morbidity (WHO 2005). Less conclusive effects were reported, however, for cardiovascular disease and the more recent studies reported no association. A number of epidemiological studies, for example, found no association between acute exposure to ambient O_3 concentration and hospital admission due to cardiovascular diseases (WHO, 2008, Anderson et al., 2005); significant associations were observed in a few studies only. Nevertheless, results of short-term effect studies do suggest a link between ozone exposures and adverse cardiovascular events such as myocardial infarction (Ruidavets et al., 2005, Mustafić et al., 2012), heart failure (Hoek et al., 2001), and life-threatening arrhythmias (Rich et al., 2005).

O_3 is not only a risk factor for increased morbidity but is also estimated to be responsible for ca. 3 million premature deaths world-wide each year, according to the World Health Organization (WHO, 2006). It is also estimated that, in the European Union (25 countries), about 21,000 premature deaths occur annually after days with high O_3 levels (WHO, 2008).

Four meta-analyses have been undertaken of the relationship between O_3 and mortality (UNECE, 2008). These suggested significant, independent associations between O_3 exposures and different causes of mortality. Impacts on respiratory mortality are strongest; those on cardiovascular mortality seem to be weaker. These effects are not influenced by other air pollutants, weather factors (e.g. temperature and humidity), season or modelling strategy (WHO, 2006). More information on individual studies relevant to this research can be found in Appendix A, Section I.

2.1.6 Long term health effects

The WHO (2008) report concluded that evidence for the long term effects of ozone has strengthened over more recent years and, while still less conclusive than short term effects, new evidence is emerging, for example on small airway function and asthma development. Epidemiological evidence of chronic effects is less conclusive than animal and autopsy studies, largely due to limitations in exposure assessment (WHO, 2008).

Recent evidence has also suggested that lung diseases in adulthood are linked to conditions occurring during development in foetal and childhood life (Narang, 2010). The foetal origins hypothesis postulates that delays in embryo growth and development during pregnancy could contribute to morbidity later in life. The effects may not only occur within childhood but can extend to adulthood, affecting metabolism, and potentially leading to heart problems or diabetes (Osmond

and Barker, 2000). Recently, therefore, many papers have explored the association between air pollution and birth complications, including low birth weight (LBW[8]), preterm birth (PTB[9]), small for gestational age (SGA[10]) and congenital anomalies (e.g. heart defects and cleft lip). Overall, there is a growing body of evidence suggesting that exposure to O_3 during pregnancy is associated with adverse birth outcomes (Salam et al, 2005; Hawang and Jaakkola, 2008; Hansen et al, 2009). Nevertheless, not all studies have reported significant associations between ambient O_3 and adverse birth outcomes, and several studies have failed to detect any effect (Lee et al., 2008, Hansen et al., 2007, Dugandzic et al., 2006). The literature to date is thus inconclusive. WHO (2008) and Derwent et al. (2008) accordingly argue that more epidemiological studies based on cohorts of susceptible individuals are necessary in order to assess and confirm the results of long-term exposure to O_3.

At least part of the variation in results reported in these studies of adverse birth outcomes might be associated with the choice of exposure metric and methodology. As noted most studies rely on data from routine monitoring sites. The distance between these and the participants may vary greatly, so how well they represent actual exposures, and the levels of uncertainty that might exist, are difficult to assess. Different ways of applying these exposure estimates to the study population have also been used. For example, one study (Salam et al., 2005) estimated exposure by weighting the data from the three nearest monitoring stations, up to a distance of 50 km from the participant's zip code, using an inverse distance squared interpolation. Exposures were estimated as monthly averages of O_3 concentrations between 10:00am and 6:00pm, and allowance was made for both temperature and elevation; in the event that the nearest monitoring station was located within 5 km from the maternal ZIP code, data from that station was used directly, instead of by interpolation. Another study restricted participants to those living at zip codes within 4 km of the three monitoring stations within the study area (Le et al., 2012).

These different ways of estimating exposures may have important effects on findings. Hansen et al. (2009), for example, reported that, when analysing all births for mothers residing within 12 km of the nearest monitoring stations, the association was not significant; however, for births within 6 km of the nearest monitoring sites there was a significant association between ambient air pollution and risk of heart defects. Notably, in the studies that found no association between ambient O_3 concentrations and birth outcomes, exposures were estimated mainly by using between one and

[8] Birth weight less than 2500g
[9] Gestational age at birth less than 37 completed weeks
[10] Birth weight below the 10th percentile of infants born at a given gestational age

four monitoring stations, regardless of the distance between the participants' residential addresses and the monitoring stations (Hansen et al., 2007, Lee et al., 2008, Dugandzic et al., 2006).

In summary comparison between the long-term health studies of O_3 is difficult due to differences in: 1) study type (e.g. retrospective, prospective, etc.); 2) participant characteristics (e.g. children, adult, and elderly); 3) exposure estimation (e.g. based directly on measurements from nearest monitoring station, interpolation, personal samplers); and 4) differences in symptoms studied. Definitive conclusions are hard to derive, although most of the studies tend to suggest an association between ambient O_3 exposure and respiratory disease.

2.1.7 Health Impact Assessment

The estimation of relative risk is a fundamental step in any health impact assessment (HIA); health risks are simply a measure of the probability of experiencing health problems in response to a defined change in exposure. The impact of a unit change in exposure on the health of any individual nevertheless depends on his/her age, sex and susceptibility. HIA thus involves comparing the health burden under different conditions: for example the current level of air pollution compared to some alternative (e.g. counterfactual) condition, or under different future conditions (e.g. a business-as-usual and alternative, policy scenario) for a specified target population.

As such, HIA requires information on three sets of factors: the air pollution concentration (and, by extension, population exposure) under each condition; the background incidence of morbidity or mortality; and the concentration-response or exposure-response function (CRF or ERF). In this way, HIA provides a means of determining whether or not current environmental hazards (such as ambient O_3) pose a problem in terms of public health (diagnostic assessments) or the potential health costs and benefits of proposed policies or other interventions (prognostic assessments) (Briggs, 2008). Both types of assessment are a means of informing and supporting risk management. Both, also, represent the means by which environmental health sciences (epidemiology and toxicology) are translated into environmental health policy and management.

The most important health outcomes associated with ambient O_3 exposure in most epidemiological studies are acute responses - in particular in terms of pulmonary function, lung inflammatory reactions, respiratory symptoms, increased medication usage, hospital admissions and, in extreme cases, death. In some cases, also, chronic health effects have been seen, notably in long-term reduction in lung function growth (Ihorst et al., 2004, Galizia and Kinney, 1999). Amongst these,

death rates are often regarded as providing the most robust and significant indicator on health impact, and are recommended by the WHO as the main indicator of effect. This is based on several considerations. First, death is well defined and relatively accurately registered; in contrast, uncertainties often occur in the diagnosis and reporting of other health outcomes. Second the burden of death associated with ambient O_3 exposure is large, and relatively easily detectable. Third, data on background death rates are available for most countries and easy to obtain. Last but not least, death is usually the dominant influence on the overall impact on health, whether in the form of disability adjusted life years (DALYs) or economic cost.

Various estimates of the environmental health burden of O_3 pollution have been made. O_3 concentrations above 70 $\mu g/m^3$ (as a daily maximum 8-hour average), were associated with approximately 21,000 premature deaths/year in the EU-25, and 14,000 additional respiratory hospital admissions annually (WHO, 2008). The latter number was expected to increase due to population ageing, since people over 65 years of age are most at risk. Other impacts may be less severe, but nevertheless affect the daily health of large populations: for example, limited activity days, increased medication usage for respiratory diseases (especially in children), and lower respiratory symptoms are estimated to account for a total of 8 to 108 million person-days of disability yearly in the EU-25 (Watkiss et al., 2005). Most recent HIAs (WHO, 2008), however, report only short-term health effects at high concentrations, and have tended to neglect effects of short-term exposures to O_3 concentration less than $70\mu g/m^3$, or of long-term exposures. They are therefore likely to under-estimate the true burden of disease attributable to O_3.

2.2 Monitoring and modelling O_3

As detailed in Section 2.1.3, ambient O_3 concentrations vary both spatially and temporally. It is therefore important to consider both the spatial and temporal components in estimating ambient O_3 concentrations. In this section an introduction and illustration of some methods for temporal and spatial modelling are discussed.

2.2.1 O_3 measurement

Measurements of air pollution have been employed to estimate exposure in epidemiological studies since the investigations of the London fog in 1952 (Ministry of Health, 1954). Methods for measuring ambient O_3 range from simple techniques, such as use of passive samplers, to sophisticated and

expensive techniques such as those used in many ground-based monitoring systems – though all are typically based on some form of absorption spectroscopy.

Passive samplers are generally used to define the background concentration and observe long term trends in average concentrations. The advantage of these is that they can be deployed in large numbers, so can be used to determine spatial variations in O_3 concentrations but only for specific (and relatively long) averaging times such as week or month.

Active monitoring sites, on other hand, continuously measure the air pollutant concentration with high temporal resolution (e.g. every 5 seconds), and typically operate in a fixed location for a long time period (years). These monitoring systems are usually established both to define air pollution distribution and pollution sources, as well as to give an alert system for the general public.

Studies using these fixed monitors are highly dependent on the spatial distribution and density of the monitoring networks. In the case of O_3, this is a serious limitation since many networks were established primarily for ecological reasons – to monitor impacts on vegetation and habitats – so that they do not necessarily reflect the distribution of human populations well. This may bias the results of epidemiological studies. Additionally, because of their cost, the networks are inevitably too sparse to provide detailed information on spatial variations in O_3 exposures.

The majority of early studies relied on data only from the nearest monitoring site(s), and assigned this to everyone in the surrounding area, especially for epidemiological studies with a large spatial scale such as those outlined in Section 2.1.4. Jerrett et al. (2005, 2009) showed that using the nearest monitoring station might lead to underestimation of the risks of mortality associated with air pollution, and especially with O_3. Assigning the same exposure score to large numbers of people may also mean that it is impossible to identify vulnerable participants who are more sensitive to O_3, masking symptoms that would otherwise be apparent at lower concentrations. As a result, epidemiological studies may be less sensitive than they would otherwise be, and health impact and risk assessments may be affected by large uncertainties.

For detailed assessment of exposures, over both time and space, some form of modelling is essential. Of interest here are models to estimate O_3 concentrations over both short and long time periods which in turn will enable the assessment of health impacts of O_3 exposure for different time periods. Models should also provide a good basis for understanding the mechanism of O_3 production, thereby helping to identify how exposures might change in response to policies (e.g. for risk assessments or HIA) and guiding the establishment of policies to reduce exposures.

Use of models to estimate an exposure is nevertheless limited by a number of factors. One is undoubtedly the lack of awareness about the models and their performance amongst epidemiologists. Another is the complexity, and severe data demands, of some of the more sophisticated models. It is therefore informative to review some of the different approaches that have been developed to model ground-level O_3 concentrations, both spatially and temporally. The following sections summarize the available methods highlighting their advantages and disadvantages. The spatial models that are discussed are broadly divided into process models, interpolation techniques, geostatistical models, and regression methods. The temporal models focus on moving average, autoregressive integrated moving average (ARIMA), and Fourier analysis. It should be noted, however, that dispersion models can be used to estimate both spatial and temporal variations.

2.2.2 Spatial modelling

The aim of spatial modelling is to produce a map (or database) of variation in concentrations across the study area. This may be represented either as discrete points, as a grid (raster) or as some form of choropleth map. It is typically constructed on the basis of measured concentrations of O_3, together with information about O_3 precursor emissions and/or the influence of other factors such as atmospheric chemistry, transport and dispersion processes. Depending on the type of model and its application, there may be the need for calibration with measured data from O_3 monitoring stations. Once the relationship is established the model can be used to predict the concentrations at unmonitored locations. The resolution of the spatial model depends on two main factors: the scale of O_3 variations (e.g. from metres in urban areas to kilometres to tens of kilometres in rural environments), and the spatial accuracy of the input data. The spatial models discussed here have been divided, generally, into four main types: process, interpolation, geostatistical, and regression models.

2.2.2.1 Dispersion models

Dispersion models are mathematical, dynamic models taking account of the key factors that affect pollution concentrations within the plume to calculate concentrations at different places. The key parameters within the model thus aim to represent the different atmospheric processes such as dispersion, chemical reactions and physical processes that control O_3 concentrations. The

performance of dispersion model is determined therefore by the accuracy of representation of the specific processes included in the dispersion model (Holmes and Morawska, 2006).

There are different types of dispersion models including box models, Gaussian, Lagrangian, and Eulerian models (Holmes and Morawska, 2006). The box model is the simplest. It assumes that the volume of air can be approximated in a form of a box and uses the assumption that air pollutants are homogeneously distributed in the box, thus averaging concentrations inside the box. Though widely used for broad-scale modelling, this approach clearly has limited utility for modelling highly localised O_3 concentrations, as variations in O_3 are known to be affected by the local changes in meteorology and emission rates of O_3 precursors.

Lagrangian and Eulerian dispersion models are similar to box models in that they also define the volume of air with an initial concentration. Concentrations are subsequently modelled as the box moves downwind, using either the coordinate system for Lagrangian models or a fixed three-dimensional grid in Eulerian models. The Gaussian approach is especially widely used in atmospheric dispersion models, partly because of its relative simplicity. It assumes that the air pollutant has a normal probability distribution both horizontally (across the plume) and with height. It is often used for predicting the dispersion of continuous air pollution plumes produced from different sources such as industrial stacks or linear sources such as roads. Gaussian models are considered more suitable for considering the local changes in the pollutant environment, as suggested by Colvile and Briggs (2000), and are widely used nested within Lagrangian and Eulerian models (Holmes and Morawska,2006).

The Community Multiscale Air Quality (CMAQ) model simulates the chemical and physical processes such as transport, deposition and chemical transformation which influence distribution of O_3 concentration (Daniel and Denise, 2006, Byun and Schere, 2004). This model is one of the recent air pollution dispersion models developed for regions with complex terrain and topography. CMAQ comprises three modelling components: a meteorological modelling system (e.g Model-3), a component for modelling for both the anthropogenic and biogenic emission, and a chemistry-transport modelling component, as described by Byun and Schere (2004). Sokhi et al. (2006) used CMAQ to predict hourly O_3 concentrations for July and August 2002 over London city. A model was run at a 9km averaging scale, and then refined to a 3 km scale, before finally being targetted at a 1km level. Model performance was evaluated by comparing the model estimates to observed concentrations from nine background urban sites. The correlation was 0.7, with a normalised mean

square error (NMSE) equal to 0.43 and 0.40 for 13-17 July and 14-18 August, respectively. Fractional Bias (FB) was also computed and indicated over-prediction.

At the continental scale, Daniel and Denise (2006) have used CMAQ to examine the spatial variability of O_3 concentrations in July 1996 over continental USA. A 36 x 36 km grid was used. They reportedly best predicted the hourly concentrations for intermediate observed concentrations (40-60 ppbv which is ~ 80-160 $\mu g/m^3$) obtained from 1110 monitoring sites. The model tended to over-estimate the lower concentrations in rural sites and under-estimate the higher concentration at urban and suburban sites, with a mean bias of 12 $\mu g/m^3$; however, when data was averaged over longer period the errors decreased. Shi et al. (2012) also used CMAQ to gain further understanding of the dispersion processes during events with high O_3 in southwest USA. Their model domain covered the southwest USA and had a grid cell size of 36 km. Model performance was determined (on the basis of independent ground O_3 measurements) by the correlation coefficient (R). For the 8-hr maximum O_3 concentration, for June–July 2006, the correlation was 0.5 between observed and modelled concentrations.

An intensive data collection campaign and simulation was performed by Gariazzo et al. (2007), using the Flexible Air quality regional Model (FARM) to predict different air pollutants, including O_3, in Rome. FARM is a 3-D Eulerian model that deals with multiphase chemistry of air pollutants and transport. A nested approach comprising three domains was adopted: a large domain covering the Italian peninsula (16km grid), an intermediate domain for central Italy (14km grid), and a target domain including Rome (1km grid). The last of these was used to capture meteorological and chemical processes. The performance of this model using observed data from three monitoring sites was evaluated on the basis of the FB and NMSE. Results gave $0.3>FB>-0.3$ and $NMSE<4$, implying good agreement between modelled and observed concentrations.

Global dispersion models are also available which attempt to simulate the chemical tropospheric processes that determine O_3 concentrations at broad scales. One such model is the global three dimensional model (GEOS-CHEM) developed by Bey et al. (2001). A crucial feature of this model is the simulation of O_3-NOx-HC chemistry, as a major factor controlling the formation of ground-level O_3. As inputs, the model requires observations of 20 meteorological factors, and transport rates for 24 chemical tracers to describe the O_3-NOx-HC chemistry, as well as photochemistry data for 80 pollutant species and more than 300 reactions, emissions (NOx, CO, CO_2) and deposition rates. All these data were used to simulate O_3 concentrations for the year 1994 on a 4° x 5° grid using the Goddard Earth Observing System (GEOS). Simulations for 1-year gave estimates of O_3

concentrations within approximately 20µg/m³ of the observed concentrations worldwide. However, the model tended to under-estimate summer season concentrations in northern mid-latitudes.

Another global dispersion model is the UK Meteorological Office tropospheric 3D chemistry-transport model (STOCHEM), described by Derwent (2001). This consists of three fields: a chemistry field, which includes 70 chemical species involved in 170 chemical and photochemical reactions; an emissions field, which includes global datasets of annual emissions of O_3 precursors; and a meteorology field, which includes data on 50,000 air parcels. EUROSTOCHEM is a regional-scale derivative of this model. It has been developed to predict ground-level O_3 at high temporal resolution and for a spatial resolution of 150km x 150km, to give estimates of the maximum hourly O_3 concentration. The model consists of the same three fields as its parent model STOCHEM, but increases the number of air parcels it can handle to 500,000 (Hayman et al., 2002). The EUROSTOCHEM model has been shown to reproduce well the maximum hourly O_3 concentrations observed daily at each site, but to be less successful at simulating the full diurnal behaviour.

As these global dispersion models clearly show, models of this type are exceedingly complex, and require a mass of input data, on atmospheric chemistry, emissions and meteorology. Running them also requires specialist expertise. These requirements inevitably limit their utility for many applications, including most health studies. Probably for this reason, they have not been widely used in epidemiological studies or for HIA.

In light of this, simpler models of O_3 concentrations are needed that can be applied to large study populations, at relatively high spatial and temporal resolution, on the basis of readily available data. In recent years, also, GIS have been increasingly used to model and map air pollution. The goal in this project is to use of GIS (and associated statistical) methods, as a means of developing a robust model of ground-level O_3 concentrations, for the purpose of exposure assessment.

2.2.2.2 Interpolation

Interpolation methods are perhaps the most well-established and basic method of modeling in GIS. They involve estimating conditions at unmonitored locations on the basis of information from surrounding, or nearby, locations. As Briggs (2005) thus shows, they can be used to construct an air pollution surface using data from the available monitoring sites.

Interpolation methods can be used to develop a spatial model for a set of observations (Y) from any number (n) of sites (S) (i.e. Y (S_i), i =1,….,n) distributed spatially over a predefined study area. These data are used to predict values for unmonitored locations Y (S_0) within the study area.

Many different methods of interpolation have been developed, including what can be termed global and proximal techniques (Burrough and McDonnel, 1998). The former attempt to fit a single (global) model to the complete data set; the latter develop models for each locality only on the basis of nearby sites. Trend surface analysis is the most widely used global interpolation method. In the case of air pollution, however, proximal methods are generally likely to be more effective, because they are more likely to capture the local variability that typically occurs, especially in urban areas.

Probably the simplest method of proximal interpolation is Thiessen tessellation. This creates polygons around each monitoring station so that each location is assigned to its nearest monitoring station. An advantage of the technique is its ease of use and computational speed, making it possible to estimate exposures for large numbers of individuals. It inherently assumes, however, that the pollution surface is flat and disjunctive – i.e. it does not vary within the area nearest to any site, but then changes abruptly at the boundary with the next monitoring site. Implicitly, this is what is assumed when exposures are assessed by assigning people to their nearest monitoring sites.

A range of more sophisticated, and realistic, methods of interpolation are available using GIS, each based on different underlying assumptions about the spatial structure of the air pollution surface. Inverse distance weighting (IDW) is perhaps the most commonly used in air pollution epidemiology (Briggs, 2005b). It is based on the principle that nearby data points provide more information about conditions at a target location than those further away. It thus weights the information from the different monitoring sites as an inverse function of distance, and then averages the weighted concentrations to give an estimated concentration at the target location. Different inverse distance functions may be used: while a linear function (1/d) can sometimes be assumed, in many studies a non-linear function (e.g. $1/d^2$) is preferred.

Two techniques for applying IDW are triangulated irregular networks and moving window techniques. Both these approaches are easy to apply, but need a dense network of monitoring stations to work effectively. They also depend on user decisions regarding both the function of distance and the window size and shape.

Interpolation from monitoring sites located within urban areas will often over-estimate the concentrations, because monitoring sites are often located at places of known or suspected high

concentrations (to detect non-compliance with policy standards). Also, interpolation should ideally be done only within the range of the measured data, since extrapolation outside this range is uncontrolled, and may lead to implausible estimates. In addition, simple interpolation techniques take little or no account of the factors that might affect concentrations, and thus warp the shape of the pollution surface, between monitoring sites (e.g. local emission sources). For these reasons, interpolation is rarely likely to provide the best estimates of concentrations, and cannot be considered reliable for predictions in different areas or time periods.

2.2.2.3 Geostatistical methods

1) Kriging

Kriging is a well-known geostatistical method that relies on a set of monitoring stations within the study area to estimate pollutant concentration and standard errors (kriging variation) at unmonitored locations (Jerrett et al., 2007). Kriging is comparable to IDW except that the weights are based not only on the distance between the measured points and the prediction location but also on the total spatial structure of the measured points. The general equation for kriging can be expressed as:

$$\hat{Z}(\boldsymbol{s}_o) = \sum_{i=1}^{N} \lambda_i \, \boldsymbol{Z}(\boldsymbol{s}_i) \qquad \textbf{Equation 2-7}$$

where: Zs_0 = the prediction location, λ_i = an unknown weight for the measured value at the i^{th} location, $Z(s_i)$ = the measured value at the i^{th} location, and N = the number of measured values.

The approach is based on the principle of regionalised variables with the assumption of spatial homogeneity, which implies three main components of spatial variation: drift, spatially correlated random variation and noise. The first of these is usually computed using some form of trend surface analysis. The two latter are modelled by computing the semivariance, representing the association between the difference in monitored concentrations and distance apart of each pair of monitoring sites. Once the model is computed, a moving window is passed across the map to estimate concentrations at the unmonitored locations within the study area, using the function derived from the semivariogram see Figure 2.3.

A semivariogram typically shows three features: the nugget, sill and range. The nugget represents any unresolved variation in measurement within distances less than the minimum separation of the

monitoring sites. The sill is the value of the semivariance at the range, which is the distance where the modelled semivariance levels out: beyond this distance there is no spatial autocorrelation.

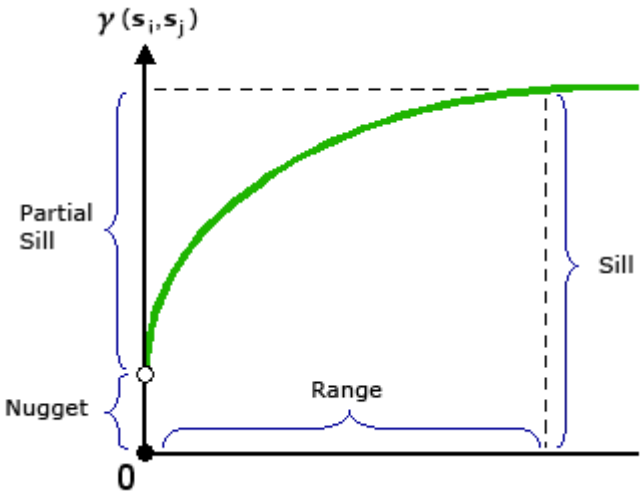


Figure 2.3: The components of the semivariogram

There are several kriging methods, of which ordinary kriging and universal kriging are the most common. The former is the most general and widely used method, and assumes that the constant mean is unknown. The latter method should only be used when there is a trend and a scientific justification can be given to describe it.

The main limitations of kriging are the requirements that monitoring sites should have a reasonably homogenous distribution and be dense enough to represent variations in pollutant concentrations at the relevant spatial scale (Jerrett et al., 2004). Any violation of this requirement will lead to errors in estimation.

2) Co-Kriging

To overcome the lack of homogeneity in data distribution, which is the case in this study, one or more secondary variables (i.e. covariates) that have a dense spatial distribution, and are correlated with the primary variable of interest, can be used in co-kriging. To get precise prediction (i.e. small error) the correlation between the secondary (predictor variables) and primary variables (dependent variable) has to be high.

Many studies have applied co-kriging for air pollution interpolation (Phillips et al., 1997, Singh et al., 2011, Sargazi et al., 2011). Philips et.al. (1997), for example, assess the impact of ambient O_3 on forest ecosystems by modelling the O_3 concentration from 165 monitoring stations in Loblolly pine in southern USA, using four interpolation methods: inverse distance weighting, inverse distance squared weighting, ordinary kriging and co-kriging. The covariate used in the last of these was the synthetic O_3 exposure potential Index which correlated with O_3 concentration at R = 0.6. The results showed that estimation of O_3 concentration was more precise when using co-kriging compared to the other three methods. Use of co-kriging is not easy with ArcGIS (the most widely used GIS), however, so specialist software has to be used. This has greatly limited its application in epidemiological studies. Care is also needed in selecting relevant, and well-measured, covariates, which must reflect the local variations effectively.

2.2.2.4 Regression based models

Land use regression (LUR) is now a popular GIS-based approach for estimating air pollution exposure for participants in epidemiological studies. LUR uses geographic attributes to predict the spatial distribution of air pollution over a defined area, typically for long-term average concentrations.

LUR (originally known as regression mapping) was developed in the Small Area Variations in Air quality and Health (SAVIAH) study undertaken by Briggs et al. (1997). LUR uses the monitored pollutant concentrations of interest as the dependent (predicted) variable, and variables such as traffic, land cover, and other geographic variables of interest (as proxies for air pollution sources) in defined distance(s) as the independent (predictor) variables, in a multivariate regression model. Pollution concentrations can then be predicted for any location, such as individual homes, using the derived parameter estimates from the regression model.

Over recent years, this approach has become widely used in epidemiological studies (Beelen et al., 2009, Beelen et al., 2008, Jerrett et al., 2007, Ryan, 2007, Sahsuvaroglu et al., 2006, Ross et al., 2005). The performance of LUR techniques nevertheless varies depending on the nature of the spatial variation and the specific characteristics of the data (Briggs, 2007) and the pollutant being modelled. In addition the variables included in LUR models have varied by study, depending on the quality and type of data available, as well as the geographic location of the study area. There is some evidence from studies that there is correlation between model performance (e.g. the coefficient of determination between predicted and observed concentrations) and the number of sample sites, but the exact location of the sites and how they reflect the spatial distribution of emissions have a

strong impact on the model R^2 (Basagaña et al., 2012, Wang et al., 2012). Several studies have shown that incorporation of site-specific variables into LUR methods enables detection of small area variations more effectively than other methods of interpolation (Gulliver et al., 2011, Gilliland et al., 2005, Jerrett et al., 2004, Collins, 1998, Briggs et al., 1997).

Generally, LUR seems to provide good estimates of air pollution, with R^2 typically in the order of 0.6-0.8, and low standard errors (Hoek et al., 2008) . Briggs et al. (2000) and de Hoogh (1999) found that LUR predicted measured concentrations better than dispersion models. Compared to dispersion modelling, LUR is also an easy procedure that is far less demanding in terms of data or computation. Jerrett et al. (2004) also showed that LUR was more effective than geostatistical techniques (kriging) and dispersion models in predicting localized variation of air pollution.

Few attempts have been made to apply any of these GIS-based techniques to O_3, especially at the continental scale. Nevertheless, the studies carried out to date suggest that spatial modelling of long-term O_3 concentrations could be achieved using GIS based techniques, especially LUR (Beelen et al., 2009, Nikiforov et al., 1998). Attempts to enhance LUR by incorporating time into these modelling techniques have also not been made to date, especially in terms of short-term (daily to seasonal) variability. Potential methods for temporal modelling are described next.

2.2.3 Temporal modelling

Like spatial variability, temporal variations in air pollution can be thought of as comprising three different elements: a systematic or repeated variation (periodicity) due to the regular effect of repeating factors, such as temperature variations between seasons, photochemical variations between day and night, or the diurnal variation in traffic volumes; a random, temporally correlated component, due to extraneous factors, such as weather; and true random variation, or noise. The question in modelling temporal variations is how to describe these three components of variation.

Modelling the random temporally correlated influences is only feasible if models can be generated of the extraneous agents causing these disturbances. In many cases, therefore, temporal models are more concerned with the systematic variations. In many time series studies, for example, variables are incorporated to reflect the seasonal variations that might impact on health. Time series data may, however, include a number of different repetitive patterns, which are partly hidden by noise. Detecting these, and determining appropriate models to represent them, is therefore difficult. Numerous approaches have been developed. Three possible approaches discussed here include

moving average (Cleveland and Devlin, 1988), autoregressive integrated moving average (ARIMA) and Fourier analysis model (Piegorsch and Bailer, 2005).

2.2.3.1 Moving average

The simplest approach to model a temporal signature is moving averages. Locally weighted regression or LOESS functions are also a robust extension of the moving average approach. Moving averaging is used with time series data to smooth the short-term fluctuations and emphasise the long-term fluctuations. LOESS was developed by Cleveland and Devlin (1988) for three purposes: to derive modelling diagnostics, to provide a nonparametric regression surface and for data exploration. This is done by building in a polynomial function using weighted least squares, where near points are weighted more than far points based on a scatterplot of the data. This method is used to add some flexibility to linear regression by incorporating a non-parametric function which then makes the linear effects easily apparent. The LOESS function is defined based on: order of polynomial *d*, weighted function *W*, number of iterations for a robust fitting *t*, and the size of the smoothing parameter *f*. The first three items have to be suited to all situations, while the smoothing parameter has to suit the data based on the scatterplot (Cleveland, 1979).

In assessing the geophysical (i.e. solar radiation and geomagnetic activities) effects on incidence of suicide, Partonen et al. (2004) analysed data for 27,469 individuals in Finland, who committed suicide between the period of 1979 to 1999. In Poisson regression (using LOESS functions) the daily number of suicides and the daily mean and maximum levels of geomagnetic activity were modelled with the population by region as the denominator. They found a strong seasonal effect, with the greatest occurrence of suicide in the spring.

In epidemiological studies exploring the impact of air pollution on human health (Schwartz, 2005, Schwartz, 1999) likewise used LOESS functions to control the nonlinear dependency of health impact with weather and season. For instance, to explore the association between air pollution and daily hospital admissions for heart disease since 1993, Schwartz (1999) used Poisson regression including nonparametric smoothing functions for the included covariates as expressed in equation 2-8:

$$\mathbf{Log\,[E(Y)] = \beta_o + S_1\,(X_1) + \ldots\ldots\ldots + S_n(X_n)} \qquad \textbf{Equation 2-8}$$

where: E(Y) is the expected value of the daily count of admission, X is the covariate, and S is the smoothing function. The result confirmed that the association between PM_{10} and CO with hospital admission was independent of weather and other pollutants (O_3 and SO_2).

Moving averages are an effective way of detecting and describing both trends and systematic (e.g. seasonal) variation in time series data. The models they provide, however, are entirely conditioned to the data used, and cannot be assumed to represent underlying variations which might be applicable more generally. For this reason, it might be expected that new models need to be developed for each data set or each situation (e.g. area, study period or policy context).

2.2.3.2 ARIMA

ARIMA methods are a further statistical extension of the moving average approach. As developed by Box and Jenkins (1976) they identify the hidden pattern in time series data for the purpose of generating forecasts. As expressed in equation 2-9, ARIMA uses the past dataset (x_{t-p}) and/or error (ε_t) to model the general dataset (x_t), where AR (p) determines the number of steps in the past dataset that are needed to forecast the present dataset. MA (q) represents the white noise up to lag (q), while Φ and Θ are the autoregressive and moving average coefficients, respectively (Makridakis and Hibon, 1997).

$$X_t=\Phi_1 x_{t-1} + \Phi_2 x_{t-2}+\ldots+ \Phi_p x_{t-p}+ \varepsilon_t - \Theta_1 \varepsilon_{t-1}- \Theta_2 \varepsilon_{t-1}-\ldots.- \Theta_q \varepsilon_{t-q} \qquad \textbf{Equation 2-9}$$

ARIMA attempts to exploit the autocorrelation for predictive purposes. Analysis, typically, involves three stages. First, based on patterns of autocorrelation and partial autocorrelation a primary model is developed. Then, parameters of the model are estimated on a temporary basis. The final stage assesses to what extent the values of the parameters estimated in the primary model are consistent. These three stages are repeated in order to achieve a model consistent with a specified time series of data.

ARIMA has been used in a range of different disciplines: for example, in predicting the number of hospital beds during a disease outbreak (Earnest et al., 2005), predicting the daily physical activity of 950 participants in the Netherlands to promote healthier lifestyle programme (Long et al., 2009), and forecasting air pollution (Kumar and Jain, 2010). In the last of these, daily concentrations of O_3, NO_2, NO and CO for one year in India were used to build a forecasting model for daily

concentrations. For 20 of the sample forecasts, one day step ahead, RMSE values of 11.4, 10.9, 15.7 and 239.3 for O_3, NO_2, NO and CO, respectively were achieved.

As the above example illustrates, ARIMA models are most effective when the temporal extrapolation (i.e. prediction) is short. Use of ARIMA models to predict over the longer term is likely to be unreliable, for there is little guarantee that the patterns identified in the past data represent underlying patterns, and will be replicated into the future. The success of an analysis using ARIMA also depends on the level of experience of the researcher (Bails and Peppers (1982). If used incorrectly, there might be a mis-specification of the periodic properties of the predicted data (Tiao and Grupe, 1980). In addition, ARIMA models may be arbitrary and difficult to interpret (El Raey et al., 2006). Based on a review of time series extrapolation, Armstrong (2001) concludes that models work effectively only if: the pattern of the time series results from a straightforward reaction (i.e. it is easy to explain the relationship of the series), dataset properties are stationary through the time, and unsystematic variation caused by random sources is not important. Unfortunately for the researcher, these three conditions rarely occur together.

2.2.3.3 Fourier analysis

Fourier analysis attempts to reveal and describe hidden systematic patterns in time series data. It does so in the form of the sum of trigonometric functions (sine and cosine). Typically, successive trigonometric function terms - referred to here as time functions - are added to produce a complex curve using regression analysis.

Fourier analysis is a well-established approach which has been widely used in contexts involving time series comprising one or more repeated signals, disturbed by noise. Much of the development of the approach has been done in engineering – especially in relation to signal processing of communications data (e.g. radio, telephone). More recently, however, the approach has found favour in other disciplines, including Technology, economics and environmental sciences.

It is widely used in technological fields that require methods to describe and model time-varying phenomena such as digital communications system (Proakis and Salehi, 2002), where the analysis of signals in the frequency domain is done through employing Fourier analysis. In the environmental field, (Richards and Baker, 2002) explored the quality of water in four rivers in North-western Ohio between 1975 and 1995. In their analysis, seasonal variation in flow over time was incorporated as a covariate in regression analysis in the form of time functions in pairs ($2.\pi.time$ and $4.\pi.time$). The

first pair produces waves with one maximum and minimum while the second pair produces two maxima and minima per year. Another study was conducted to analyse the seasonal and inter-annual variation in satellite time series imagery for different types of land cover condition using Fourier analysis. Performance was compared against a normalized difference vegetation index, and the results confirmed the ability of Fourier analysis to characterise and identify the different type of land cover conditions occurring in southwest Kansas, USA (Jakubauskas et al., 2002).

In epidemiological studies, time functions have been used for several different purposes. A study by Harlap (1974) exploring the role of environmental factors in the incidence of Down's syndrome presented the monthly rate of Down's incidences adjusted by maternal age for 42,340 births in 1964-1970 in terms of time functions for different numbers of maxima and minima during the year (1.π.time, 2.π.time, 3.π.time, and 4.π.time). These functions were then offered in a regression analysis, which confirmed the existence of a six-month cycle with maxima in spring and autumn. Another study explored the impacts of air pollution on daily mortally from respiratory and cardiovascular diseases in Hong Kong from 1995-1998. This study used a regression model including day of the time series, days of the week, meteorological factors, and time functions to represent the seasonality variations (Wong et al., 2002).

Skene et al. (2010) estimated daily NO_2 concentrations in Connecticut using variables representing traffic volume, land use, population density and altitude with seasonal variation of NO_2 included as covariates in the regression analysis. Adjustment for seasonality is represented by two approaches: 1) a spline function (type of moving average) and 2) time functions. The authors preferred the latter approach as it was found to be more generalized and was not tied to the specific time period of the data used in the study.

In a forecasting study, Damsleth and Spjotvoll (1982) developed a model using time functions to represent long-term predictions for the sunspot series. In a stepwise approach, a large number of time functions were used in series of pairs of sine and cosine with sequence frequencies from 2 to 8 in Fourier analysis. The analysis was terminated when there were no additional significant time functions. They also compared the approach with ARIMA, and reported that the sunspot series was predicted better using Fourier analysis.

As the Fourier analysis model can only present the systematic variation, other covariates representing the unsystematic (non-periodic) variation need to be considered in order to reduce the white noise (i.e. residual error). Much of this non-systematic variation in O_3 concentrations may be

expected to be due to meteorological factors (see Section 2.1.1). These might therefore usefully be incorporated into the model, by regression analysis or other techniques.

In a study by Vingarzan and Taylor (2003), for example, the average daily maximum O_3 concentration, both annually and in summer seasons, was modelled using multiple regression. This model included temporal cycles (from 7 days to 5 years) to represent the seasonal trend, along with meteorological factors, regressed against the O_3 daily maxima for 5 year intervals across the period from 1985-2000. Daily meteorological factors (average wind speed, maximum temperature, and sun duration hours) were included to adjust for the local variation in O_3 or its precursors. Lou Thompson et al. (2001), in a review study, reported that meteorological adjustment of O_3 has been widely used in estimating O_3 time trends using regression analysis. They also noted that the most commonly used meteorological factors were temperature, wind speed, solar radiation and total precipitation.

2.3 Rationale of the selected modelling approach

In recent years, a number of studies have tried to build space–time models for air pollution, typically to provide precise exposure measurement for epidemiological studies or health risk assessment. In doing so, they typically consider three key requirements:

1) they try to cover a large study area in order to ensure significant contrasts in exposure and health risks between different countries (or states) as well as to achieve sufficient statistical power for detecting infrequent health outcomes;

2) they try to assess exposures at a local or individual scale to take account of within city variations in pollution concentrations; and

3) they try to allow for temporal, as well as spatial, variations in pollutant concentrations, in order to take account of the different exposure experiences of people at different times (e.g. cohorts of children who are born at different times of the year).

In principle, the most effective means of meeting these requirements is through the use of an empirically validated dispersion model. For this to work, however, not only the model but also all the relevant input data needs to exist for the study area and time period of interest. Both the model and the data also have to be at a suitable level of spatial and temporal aggregation.

As has been indicated, these conditions are not met in the case of O_3: a usable dispersion model, able to provide estimates of concentrations at a high spatial resolution (<1km) and temporal resolution (ca. daily) over large study areas, with the available data and processing capacity, is not available. Alternative methods have to be sought.

As Briggs et al. (2000, 2005) suggest, one alternative is to develop temporally varying land use regression models. In principle, this is possible as long as suitable time-varying predictors are available. In some instances, this is feasible – e.g. where dynamic (i.e. regularly updated) emission inventories exist that indicate the time-varying patterns of source intensity across the study area. Often, however, and especially for short-term variations, such data are not available. In these cases, unless process models can be used, different modelling approaches need to be combined to model the space-time patterns of pollution. This is the case here.

Using a general additive mixed model, Yanosky et al. (2008) modelled the monthly space-time trend of PM_{10} concentration in 13 states in the North-eastern and mid-western USA between 1988-2002. Monthly data from 922 monitoring sites were used in a two stage analysis, using both GIS-derived vector covariates (constant with time) and local meteorological factors and area-source emissions (varying with time). In the first stage, the site-specific term was modelled adjusting for time-varying covariates. The second stage modelled the time-invariant site-specific term, using the GIS-derived variables.

Another two stage model was developed by Dadvand et al. (2011) for estimation of weekly concentrations of black smoke between 1985-1996 at residential postcodes across northeast England. This model was developed using four libraries of the R statistical package. In this study black smoke data were obtained from 56 non-automatic monitoring sites with sparse daily concentration measurements. Meteorological factors (temperature, precipitation and wind speed) from a number of monitoring stations across the region and GIS-derived covariates such as traffic, land cover classes, and industrial activity were also available. The first stage modelled the temporal trend for the whole region on the basis of the average BS concentration from all monitoring stations for each of the 627 weeks, using a dynamic model representing the seasonal variation in BS concentration as a function of meteorological factors. This produced an offset to be used in the second stage. The second stage modelled the spatial variation at all locations using the GIS-derived covariates in a linear regression including the first stage result. Thus, two different models were selected, and their results combined, to represent the spatial and temporal components of variation.

The same two stage models cannot easily be applied across Europe for a number of reasons. The foremost is the size of the data matrix (hourly O_3 concentrations multiply by 1211 monitoring station distributed over Western Europe). The second constraint is the limited availability of regional meteorological data on an hourly basis, which are key covariates for modelling the temporal variability. Kyriakidis and Journel (1999) also argue that a stochastic model cannot explain the variations in concentrations produced via physical processes unless a deterministic model is included. Nevertheless, a similar two-stage analysis (one to model the spatial variations and the other to model temporal variations) does seem appropriate. The general model of space-time O_3 data measured at Y(x,t) in continuous time (t) and space (x) can thus be represented as:

$$\mathbf{Y(x,t)= \mu (x,t) + Y_i (x,t)} \qquad \textbf{Equation 2-10}$$

In this, the spatial component of the model is denoted by $\mu(x,t)$, where t represents the trend (mean) over the whole time period (here, six years), which depends on location (x). The temporal component Y_i (x,t) is the fluctuation around this trend in both space and time.

The spatial component μ (x,t) will be calculated here using a stochastic model (LUR). This was selected because, as described in Section 2.2.2.4, it has become one of the most popular GIS-based approaches to estimating and mapping pollutant concentrations, and has proved to be both practicable and reliable across a range of spatial scales.

The temporal component Y_i (x,t) will be calculated using a semi-deterministic model. For this purpose, Fourier analysis (Section 2.2.3.3) is used, to represent systematic variations in O_3 concentrations over different time periods (hourly, daily and seasonally), and meteorological factors (MFs) are then incorporated to explain the non-systematic (weather-related) variations. This approach is defined as semi-deterministic in that the time functions used to describe the systematic variations are based on a priori expectations about the shape of the patterns in O_3 concentrations over the different time periods, but these are then calibrated to the measured data using stochastic (regression) methods. Likewise, regression analysis is used to select and weight the relevant MF.

This methodology is outlined in Figure 2.6. As this shows, there are five main phases in the analysis. The first phase involves classifying all monitoring sites in the study area into site types, on the basis of indicators describing the temporal variations in O_3. The relationship between these site types and selected environmental variables (e.g. land cover, topography, climate) is then determined, using

MLOR to estimate the probability of membership (P) of all site types across a 100 metre grid. This will be addressed in Chapter 4.

In the second phase of modelling (Chapter 5), a spatial model was developed to estimate the long term mean O_3 concentration at all locations across the study area (i.e. for a 100 metre grid). For this purpose, a LUR model was built using environmental variables as the predictors, and measured mean O_3 concentrations for the training dataset of monitoring sites across Western Europe as the dependent variable.

In the third phase, Fourier analysis was used to develop a model of the systematic temporal variability of O_3 concentrations in each site type (Chapter 6). This was done by creating a series of basic time (sine and cosine) functions to represent the expected patterns of variation in O_3 concentrations for different time periods (seasonal, weeks, days). These were then used as independent variables in a regression analysis against the deviation (offset) from the long-term mean O_3 concentration, at each monitoring site, to develop a series of TMs – one for each site type. The resulting models were then weighted using the probabilities of site-type membership (from phase 1) to estimate time-varying concentrations at each location.

In the fourth phase of modelling (Chapter 7) the spatial and weighted temporal models (WTM) (from phases 2 and 3) were combined for each site-type. This provides what is termed the base model:

Base space-time Model = LUR + [P1.WTM1+P2.WTM2+P3.WTM3…Pn.TMn].

The base model was developed and applied across Western Europe.

Finally, in phase 5, MFs were incorporated to take account of non-systematic temporal variations. The full model is thus defined as:

Full space-time Model = a.[Base model] + [b1.MF1+b2.MF2…bn.MFn]

This was developed and applied for two study areas – Rome and the Netherlands.

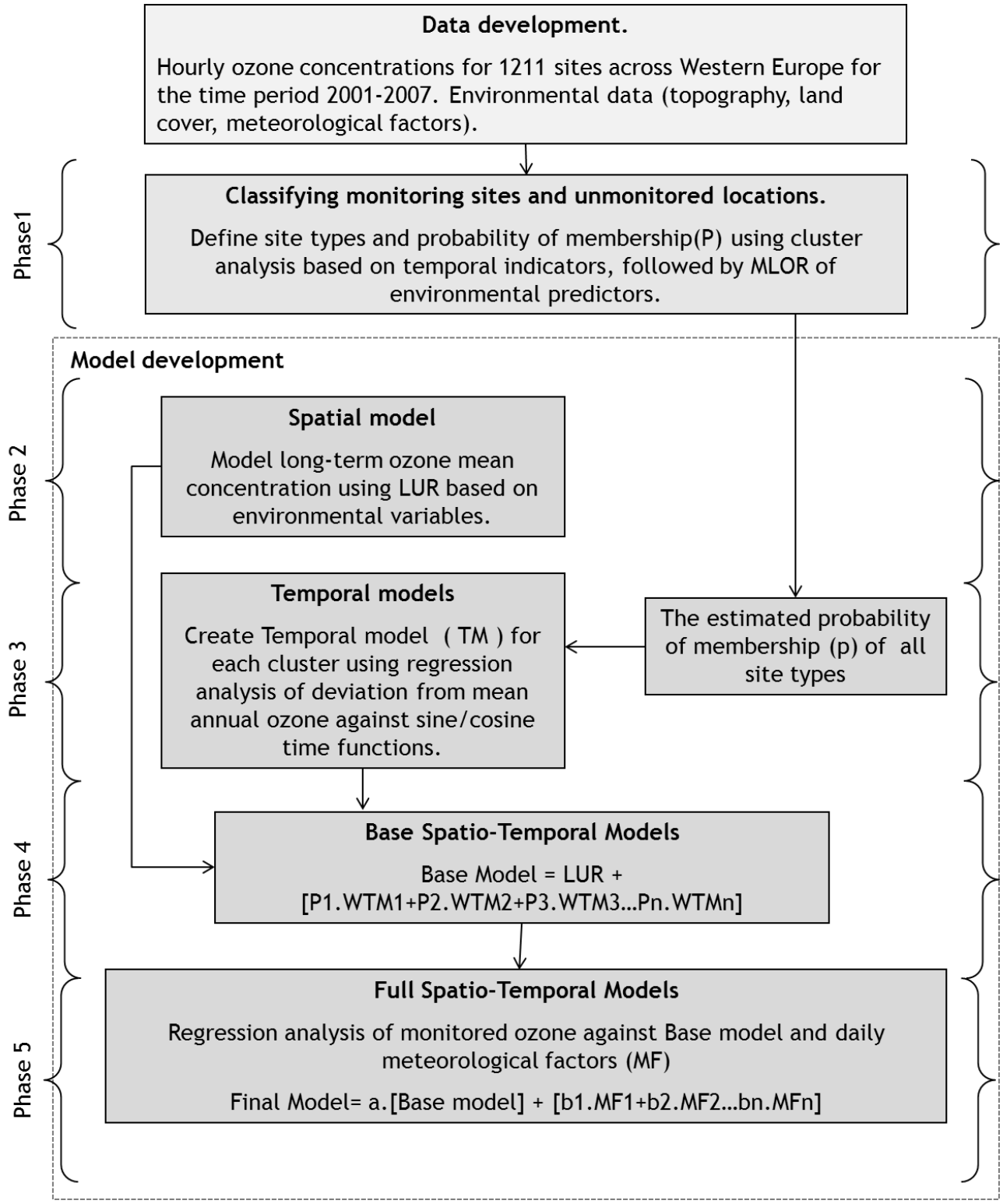


Figure 2.4 The five phases of modelling

3 Data collection and pre-processing

3.1 Study area and period

O_3 and some of its precursors are known to travel long distances. They thus represent a trans-boundary issue. Modelling likewise needs to be done across large areas, not only to reflect these trans-boundary impacts on health, but also to enable exposure assessments in epidemiological studies covering populations large enough to detect the effects of O_3 against the background of other risk factors. Likewise, research needs to cover a long enough period to represent average conditions within this study area, and to identify and take account of any temporal trends of variations that might occur. At the same time, for research purposes, it is important to ensure that all the relevant data are available as readily as possible, and at a suitable quality. Otherwise, unnecessary time is spent trying to source, acquire, check and correct data and/or results of the research are made suspect by uncertainties in the data.

The study area used in this study therefore encompasses the twelve European countries (i.e. Western Europe) – namely, Austria, Belgium, Denmark, France, Germany, Luxemburg, Ireland, Italy, the Netherlands, Portugal, Spain and the United Kingdom, as shown in Figure 3.1. These countries were chosen for a number of reasons:

1. They provide a range of contrasting environmental conditions, and O_3 concentrations, typical of that of much of the temperate world;
2. They are all subject to common air quality policies, so that policy differences are likely to have had only a small effect on O_3 concentrations;
3. They are generally rich in relevant geographic and environmental data, and these are usually relatively freely accessible;
4. As members of the EU, they are obliged to collect and provide many of the data needed in this research (notably measurements of O_3 concentrations and land use data to consistent standards) and in the same form.

The study period chosen for this research is six years, from 1st March 2001 to February 28th 2007. This period was chosen both to provide a sufficiently long time series of data and to cover the period when (at the start of the research) data availability was at its optimum. At the same time, it results in data volumes which are manageable given the computing resources available. The study period is

variable in terms of is meteorology and O_3 concentrations. It includes years both affected and unaffected by major meteorological anomalies (e.g. heat waves), and also a census year in most EU countries. This has the added advantage of allowing potential users to link the results with cotemporaneous population data for risk assessment purposes. The exact period, from March 2001 to February 2007, was selected to avoid truncating the main winter and summer seasons, which respectively mark the nadir and peak of O_3 concentrations.

Figure 3.1 Study area: Western Europe shown in yellow

3.2 GIS development

As detailed in the previous chapter, O_3 is affected by various factors yielding variations of concentration both in space and time. The relative proportions of these two sources of variation may change markedly, depending on the study area, the spatial scale of analysis, and the time

period and level of temporal aggregation. The amounts of variation are therefore not inherent properties of the real world, but artefacts of the study design.

The spatial component of variation can be expected to represent the influence of a number of environmental factors, reflecting the sources of precursor emissions, and the effects of physical transport (dispersion) and chemical reactions in the atmosphere. Modelling thus needs to be based on data relating to these various contributory factors, and the data needed both to develop the models and to run them needs to be gathered and brought together in a consistent form. Models also need to be calibrated and validated against measured observations, if they are to be considered reliable. In this thesis, therefore, data were needed on a wide range of environmental variables, that would subsequently act as predictors in the models, and for monitored O_3 concentrations at a representative range of monitoring sites. To facilitate data integration, and to provide the tools necessary for data processing, all these data were brought together in a GIS, covering the study area.

3.2.1 Spatial data development

A GIS provides a computerized analysis and mapping system, comprising all the devices needed to capture, display and integrate data in a spatial form, as demonstrated in Figure 3.2.

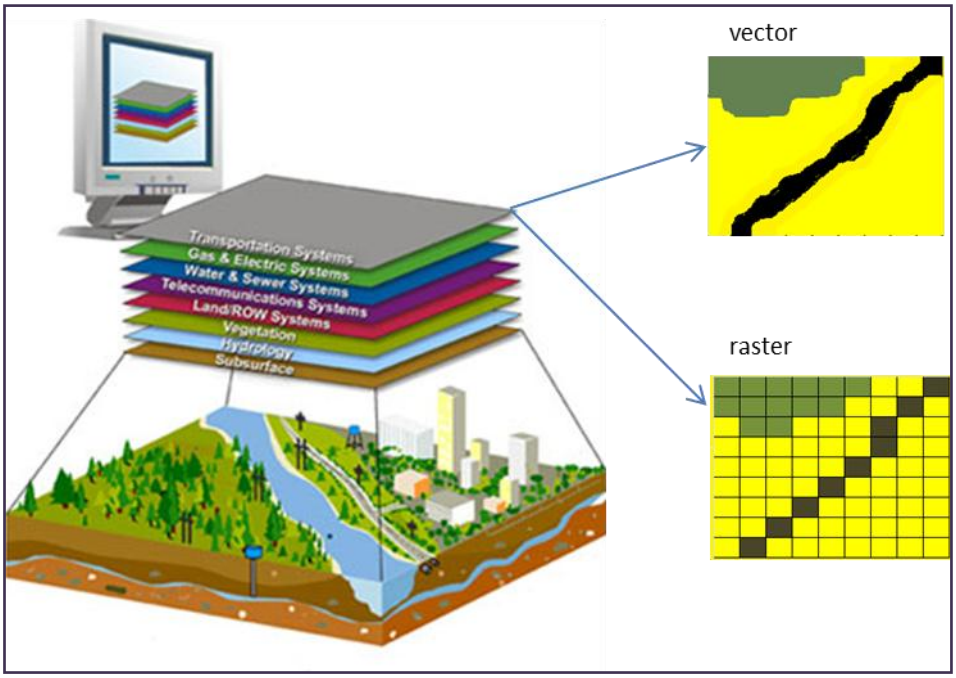


Figure 3.2 Conceptual diagram of a GIS showing thematic layers of information stored as vector or raster data

The left image from headsonfire.org

The GIS thus serves as a platform on which both to model spatial patterns of air pollution and then to compare these with measured concentrations in order to validate the models. By adding in, also, population data, either at a group or individual level, GIS also provides a powerful tool for exposure assessment (Briggs, 2005a) .

To enable exposure assessment within a GIS framework, however, all the data need to be defined in term of their geographic coordinates (i.e. geo-referenced) (Briggs, 2005b). Geo-referencing may be done in different ways, depending on the nature of the data. Monitoring sites are usually specified by their point locations, in terms of latitude and longitude (or x, y coordinates relative to a national grid). Mostly, population and health data are represented either as areas or as point locations (area centroids) of polygons representing census or other administrative districts. Road and traffic data, on the other hand, relate to lines, representing the road networks.

One of the great advantages of GIS is that they enable the rapid and easy inter-conversion and linkage of all these different data structures within a consistent framework (Briggs, 2007). This not only facilitates analysis and mapping, but also helps to reveal errors and inconsistencies in the data that would otherwise not be noticed. It also helps to avoid topological fallacies in the analysis, due to misrepresentation of the spatial relationships between different features.

To represent a geographic object in GIS (e.g. a building or a tree or a road), a form of data representation has first to be established. The two possibilities for representing a geographic object in a GIS are raster and vector. Vector systems represent features as polygons, lines, and points; raster systems divide the geographic area into regular grid-cells identified by row and column, as shown in Figure 3.2. Thus, a transportation network can be stored as lines, where the length and structure is representative of the real world roads, or the roads can be partitioned into grid cells, with the length within each segment stored as the attribute for each cell.

Using raster systems often has the advantage of greatly reducing computation times, though it obviously involves some small loss of precision, the amount of which depends on the size of the grid cells (granularity). The size of the grid cell is usually determined according to the accuracy and the resolution that is needed – i.e. based on how space needs to be conceptualised within the particular study area, and the sorts of analyses that need to be done. For example, computational modelling is generally easier in raster; topological analyses (i.e. to do with adjacency or real distance) are often better in vector. Using a raster system also greatly facilitates the combination of data and computational analysis.

In this particular work, using raster has the added advantage of enabling the efficient analysis of the large datasets needed to cover the large area, and the range of different environmental phenomena, at an appropriate spatial scale. For environmental epidemiological studies at European scale, this is certainly an attractive attribute.

From the outset, therefore, it was decided to convert the data into a raster format. All predictor variables were computed for 100m regular grid cells to produce a consistent and high resolution GIS database matching the fine resolution of the available land cover data. The Lambert Azimuthal Equal Area projection (LAEA) was the selected projection for the GIS database as it is the projection of the land cover data.

ArcGIS v10 was used to store, analyse, and manipulate the GIS data required for this study. The data were stored as ArcMap grids and exported as comma separated (*.csv) or dbase format (*.dbf) files for use in SPSS (v15 & v20) during statistical modelling.

3.2.2 O_3 concentrations data

Data on O_3 concentrations are required both to help develop and calibrate the models, and then to test their validity. These data need to be representative of the study area and period, to be consistent, and to be as free as possible of measurement error.

Measured data on O_3 concentrations for the study area countries were obtained from the AIRBASE database. This database is maintained on behalf of the EEA, and draws together measurements from routine air pollution monitoring carried out by the EU member states, under obligations imposed by EU air quality directives. As mentioned in the draft final report about AIRBASE (Spangl et al., 2007), the European Topic Centre on Air Quality (ETC_AQ) is responsible for developing and maintaining both a European Air Quality Monitoring Network (EUROAIRTNET) and an air quality information system (AIRBASE database), in close collaboration with the European countries, to provide good information to support the work of the EEA. The purpose of this collaboration is to enable and produce air quality assessments on the European scale through adequate information on air quality.

AIRBASE is developed on the basis of this EUROAIRNET network. AIRBASE has been established, improved over time and made available on the Internet under the EC Exchange of information (EoI) Decision 97/101/EC[11] in compliance with the EU air quality directives.

The objectives of the monitoring activities of the EU countries, as outlined in (Larsson, 1999), are:

- Monitoring of Member State compliance with the directives.
- Representative air quality surveillance monitoring to describe the state of and trends in air quality across Europe.
- To facilitate exposure/damage assessment with regards to health, vegetation, and materials.
- Public availability of monitoring data on-line, to inform and warn citizens, as well as to enable and inform short term abatement actions.
- Operational monitoring near specific sources to prevent undesirable pollution burdens on neighbouring areas.
- Monitoring programmes to support scientific research.

The selection criteria defining the specific areas to be monitored are intended to be representative either across the whole of Europe, or for separate regions of Europe to provide adequate spatial coverage of the air pollution situation. AIRBASE, however, comprises existing monitoring networks run at local, regional or national level, and many smaller geographic areas will not necessarily have a sufficient number of monitoring networks or sites to contribute.
Measured data on hourly O_3 concentrations for the countries of interest were obtained for the whole study period, from March 1st 2001 to February 28th 2007. Version 4 of the AIRBASE data products on the EEA data service website[12] was used for this purpose.

O_3 concentrations reported in AIRBASE are continuously monitored using automatic UV absorption analysers, and in broad terms can thus be considered consistent in terms of measurement technique. For the 6-year time period time of this study, monitoring was available for a total of 1,463 sites across the study area. To ensure that data provided a reliable basis for modelling, several criteria were used to select from these 1,463 monitoring sites:

[11] http://acm.eionet.europa.eu/databases/airbase/

[12] http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=1029

- Sites should be located within the borders of the correct country as indicated by the metadata, Figure 3.3.

- Monitored data should be available for 4 years or more from the study period, and with measurements for at least 75% of the time in any monitoring period (daily and seasonally). Based on the EU regulations this equates to 18 hours during the day, and for 22 days of each month. For this purpose, two main seasons were defined: season 1 was the three summer months and season 2 was the three winter months. The histogram in Figure 3.4 illustrates the number of sites for each country. For example approximately 210 monitoring sites in France and German have six full years of data, while 60 and 40 monitoring sites, respectively, have 5 years of full data.

- The remote islands of Spain and Italy were excluded (e.g. Mallorca, Ibiza, Corsica, and Sardinia) because they are geographically far from mainland Europe, and thus affected by different meteorological conditions and air mass regimes.

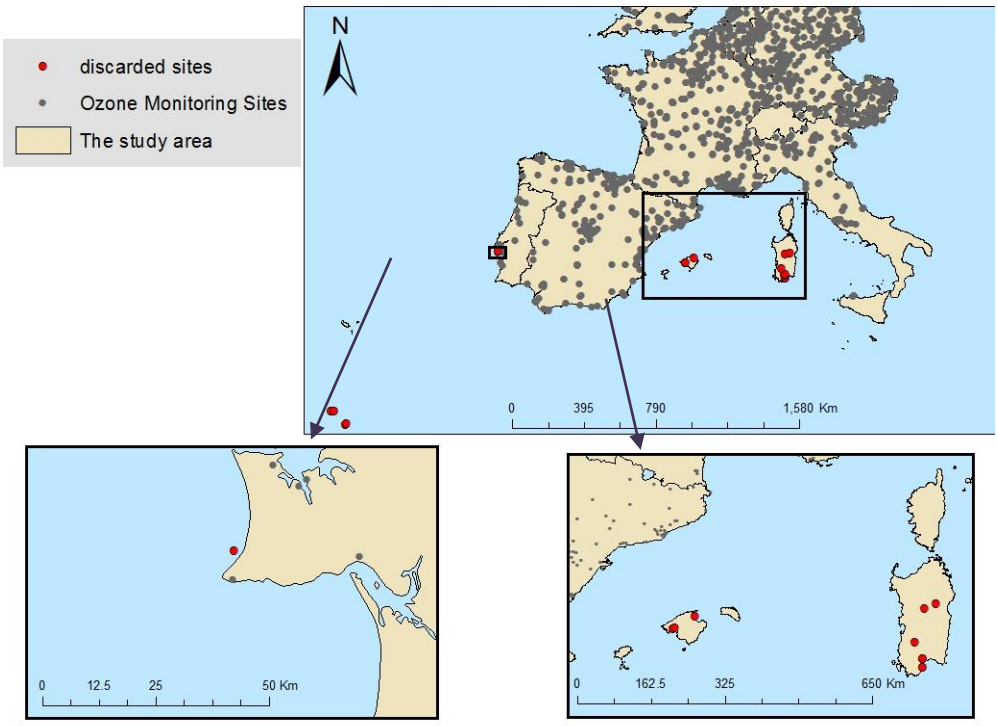


Figure 3.3 The study monitoring sites (grey dots) and discarded sites (red dots)

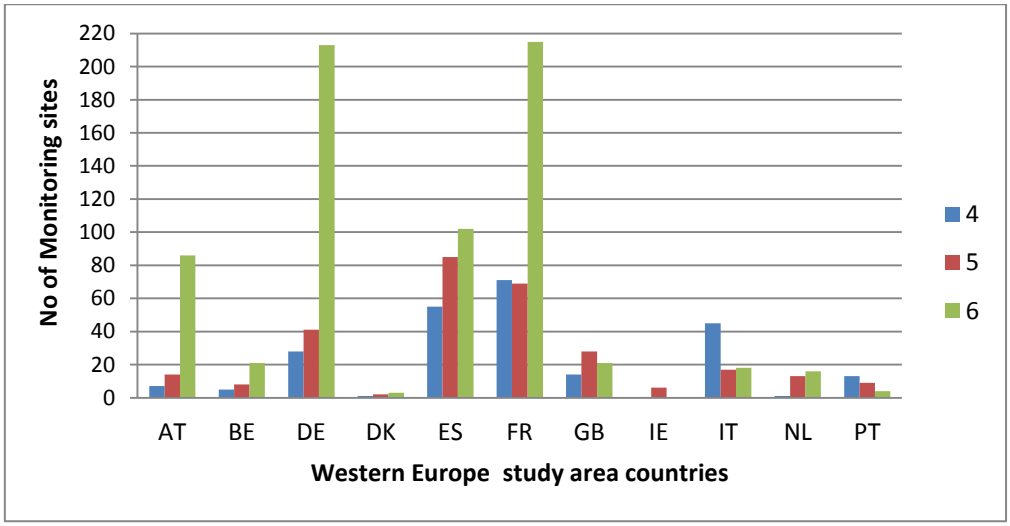


Figure 3.4 Frequency of sites, by country, with 75% hourly data capture for 4 years and more

The hourly concentration data were cleaned by recoding hourly missing or negative values to 999. Sites with less than 75% data capture were also discarded. Sites meeting the other selection criteria were identified and maintained, reducing the overall number of sites to 1,211 as shown in Figure 3.5. The full descriptive statistics for these 1,211 sites are shown in Table 3.1.

Table 3.1 Descriptive statistics for the long-term concentration (hourly data from 1st March 2001 to 28th February 2007) at the 1,211 ozone monitoring sites

country	No of sites (1211) [a]	No. of measurement (54,685,277)[b]	Min	Max	Mean	SD
Austria (AT)	107	5,067,880	0.00	336.20	58.29	36.68
Belgium (BE)	34	1,477,901	0.00	296.00	43.35	32.44
Germany (DE)	282	13,396,674	0.00	334.00	48.46	33.78
Denmark (DK)	6	267,315	0.00	195.10	47.34	26.21
Spain (ES)	232	10,154,233	0.00	470.00	48.56	32.98
France (FR)	354	16,210,237	0.00	417.00	50.85	34.31
Great Britain (GB)	63	27,03,792	0.00	327.00	42.28	26.95
Ireland (IE)	6	284,068	0.00	207.80	60.24	21.97
Italy (IT)	72	2,755,753	0.00	451.00	51.62	41.80
Netherlands (NL)	30	1,373,438	0.00	276.49	40.36	28.91
Portugal (PT)	25	993,986	0.00	358.00	48.73	30.69

a,b the total number

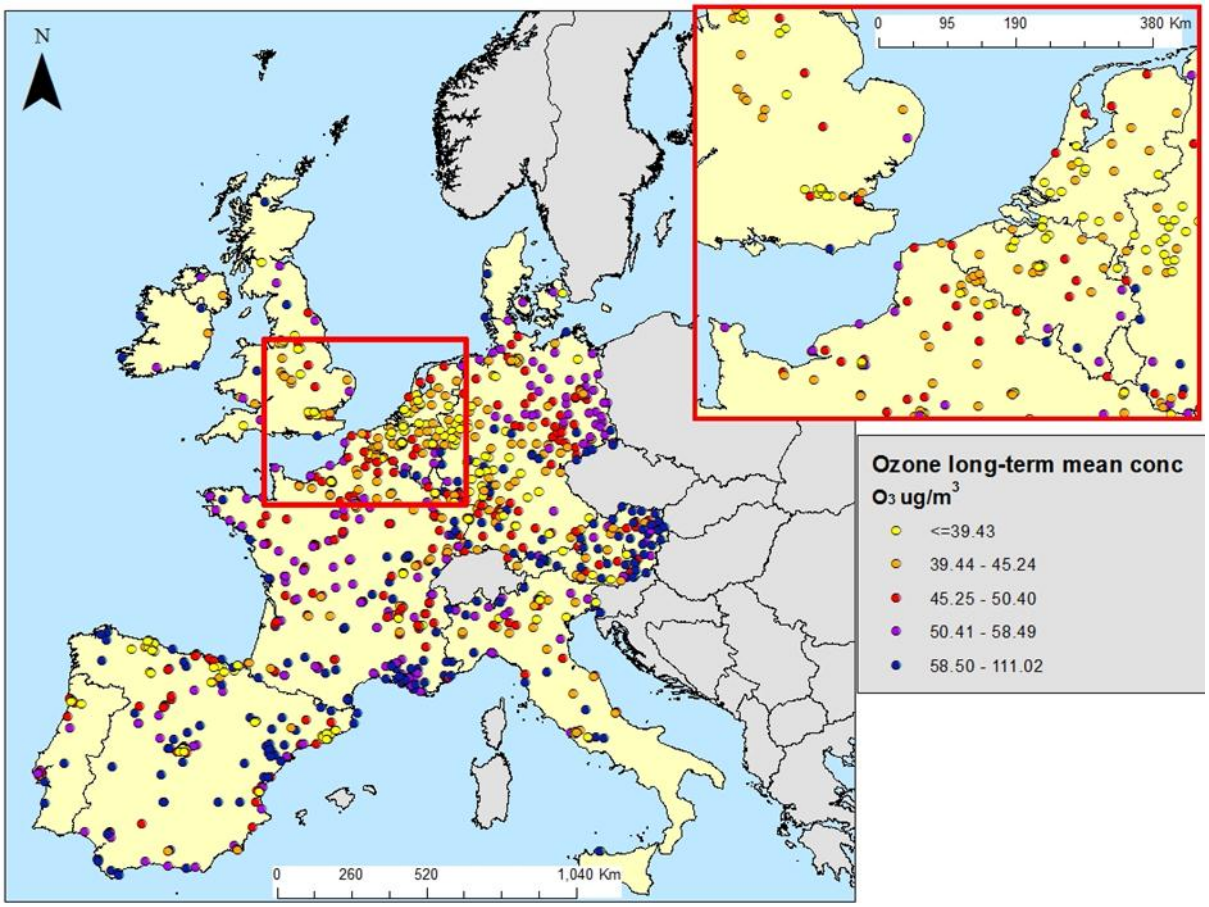


Figure 3.5 Station locations in the AIRBASE data set and the mean O_3 concentrations (6 year average)

These sites are not evenly distributed across the study area, as indicated in Figure 3.5 (map of the final selected sites). This inequality is a problem that must be recognized and addressed in the modelling to be carried out in this study, and its implications are discussed in Chapter 8.

3.2.3 Predictor variables

The effectiveness of GIS-based analysis relies on the use of relevant and accurate data on variables that can help to predict spatial variations in the phenomena under consideration, in this case O_3 concentrations. Deriving these potential predictor datasets is often time consuming, for most data need to be carefully checked, and subjected to a wide range of corrections and enhancements. Also, for prediction and mapping purposes these data need to be available (continuously) across the whole study area, including the unmonitored places (i.e. grid cells). Important predictor variables for

O_3 were identified a priori and included roads, land cover, topography, meteorological factors, and the distance to the sea; data are listed in Table 3.2.

Storing the data in the raster form makes handling and analysing easier than the vector format when the goal is to produce a concentration surface. Rasters typically contain only one single attribute; thus numerous rasters are needed to represent all the attributes of interest. Steps to create each of these are described in the subsequent sections. For some analyses, data need to be obtained for a window around a target cell – e.g. to represent the influence of emissions in surrounding areas. With vector data, these are usually extracted using buffering techniques.

With raster data, the equivalent procedure involves using a circular moving window of appropriate radius, along with relevant focal functions (sum, mean, range, and standard deviation). FOCALSUM, for example, is a moving window analysis whereby the sum of the values in a specific neighbourhood (i.e. window or cells surrounding the focal cell) is computed. The researcher specifies the size and shape of window.

Different window sizes were selected to represent the spatial zone of influence of different predictor variables. To represent local effects, windows ranging from 100 to 1000m were used; to reflect regional influences, window sizes of 5000 and 10000m were specified. As demonstrated in Figure 3.6, FOCALSUM for the 100m window consists of a neighbourhood of five grid cells, for which the values are added and the result allocated to the focal (i.e. centre) cell. This neighbourhood is passed, cell-by-cell, across the grid and the calculation repeated until the last grid cell of the study area is computed. Table 3.3 shows the FOCALSUMs used for each distance band.

After creating the original 100m grid for each of the relevant predictor variables, Model Builder in ArcGIS was used to construct the other window (or neighbourhood) averages and create the variables as a grid. These were then "intersected" with the monitoring sites using the equivalent tool for rasters, *Extract values to points*. As shown in Figure 3.7, the sequence of modelling was thus to run FOCALSUM statistics, extract values to points, update fields and export the results to DBF files. Figure 3.7 illustrates these steps to obtain land cover data for different window sizes; the same models were used to obtain all other GIS predictors simply by changing the references to dataset and folders.

Table 3.2 Overview of the predictor variables

Predictor	GIS dataset	Predictor variable	Abbreviation	Purpose	Unit	Source and resolution
Land cover variables	CORINE	High density residential land	Highdr	Scavenge O_3	Percentage	CORINE land cover 100m grid- Version 13/2000 (CLC2000) from the EEA/Resolution (100m)
	CORINE	Low density residential land	Lowdr	Scavenge O_3	Percentage	
	CORINE	Industrial and commercial land	Ind/Com	Source of O_3 precursors	Percentage	
	CORINE	Herbaceous land	Herb	Source of O_3 precursors	Percentage	
	CORINE	Agriculture land	Agri	Source of O_3 precursors Depletion of O_3	Percentage	
	CORINE	Forest land	Forst	Source of O_3 precursors (BVOC)	Percentage	
	CORINE	Open Space	Opsp	-	Percentage	
Topographical variables	CORINE	Distance to sea	D2S	Increase O_3	kilometre	Derived from the distance between each grid and the coast line
	Altitude	Altitude (height above sea level)	Alt	Increase O_3	metre	SRTM 90m Digital Elevation v4.1 produced by NASA/Resolution (90m)
	Altitude	Topex	Topex	Decrease or Increase O_3	metre	Height difference between 100m window and the mean of the surrounding 2000m cell centroids
Road length variables	road network	Motorways	MR	Scavenge O_3	metre	Eurostreets version 3.1 is a 1:10,000 digital road network
	road network	Secondary Roads	SR	Scavenge O_3	metre	
	road network	Local Roads	LR	Scavenge O_3	metre	
Meteorological factors	NETCDF	Surface solar radiation	SSR	Increase O_3	w/s	The European Commission Joint Research Centre (JRC)
	NETCDF	Total precipitation	TP	Depletion of O_3	Mm	ERA Interim, monthly means of daily means derived from ECMWF/ resolution (40km)
	NETCDF	Temperature	TMP	Increase O_3	C^o	
	NETCDF	Wind speed	WS	Decrease or Increase O_3	m/s	

Table 3.3 Window specifications based on grid cells (using FOCALSUM)

Window size (radius in m)	FOCALSUM "Circle" Distance (equivalent radius in cells)	Total number of grid cells within window
100m	1	5
300m	3	29
500m	5	81
1000m	10	317
5000m	50	7845
10000m	100	31417

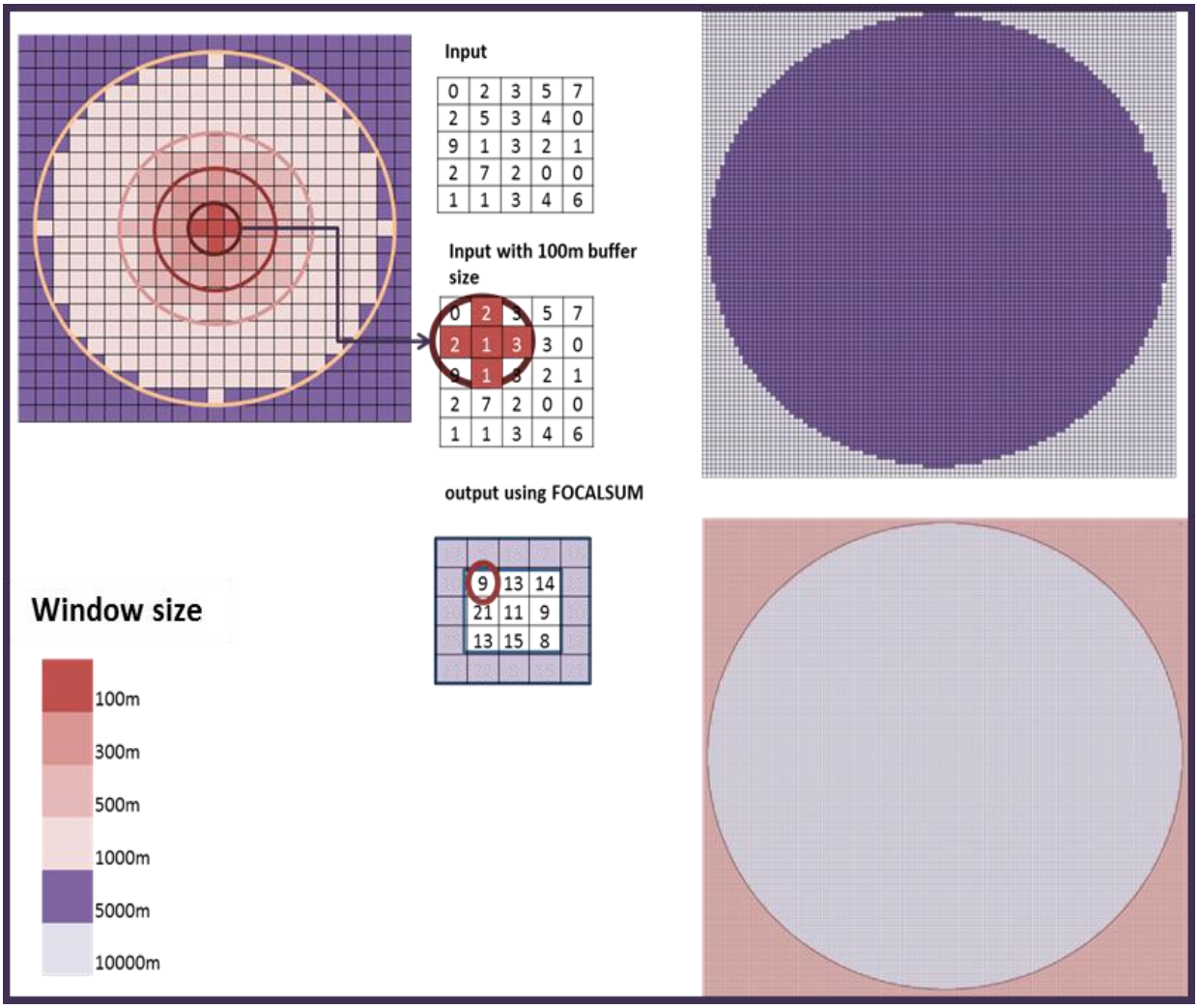


Figure 3.6 Different window sizes used in FOCALSUM for a 100m grid

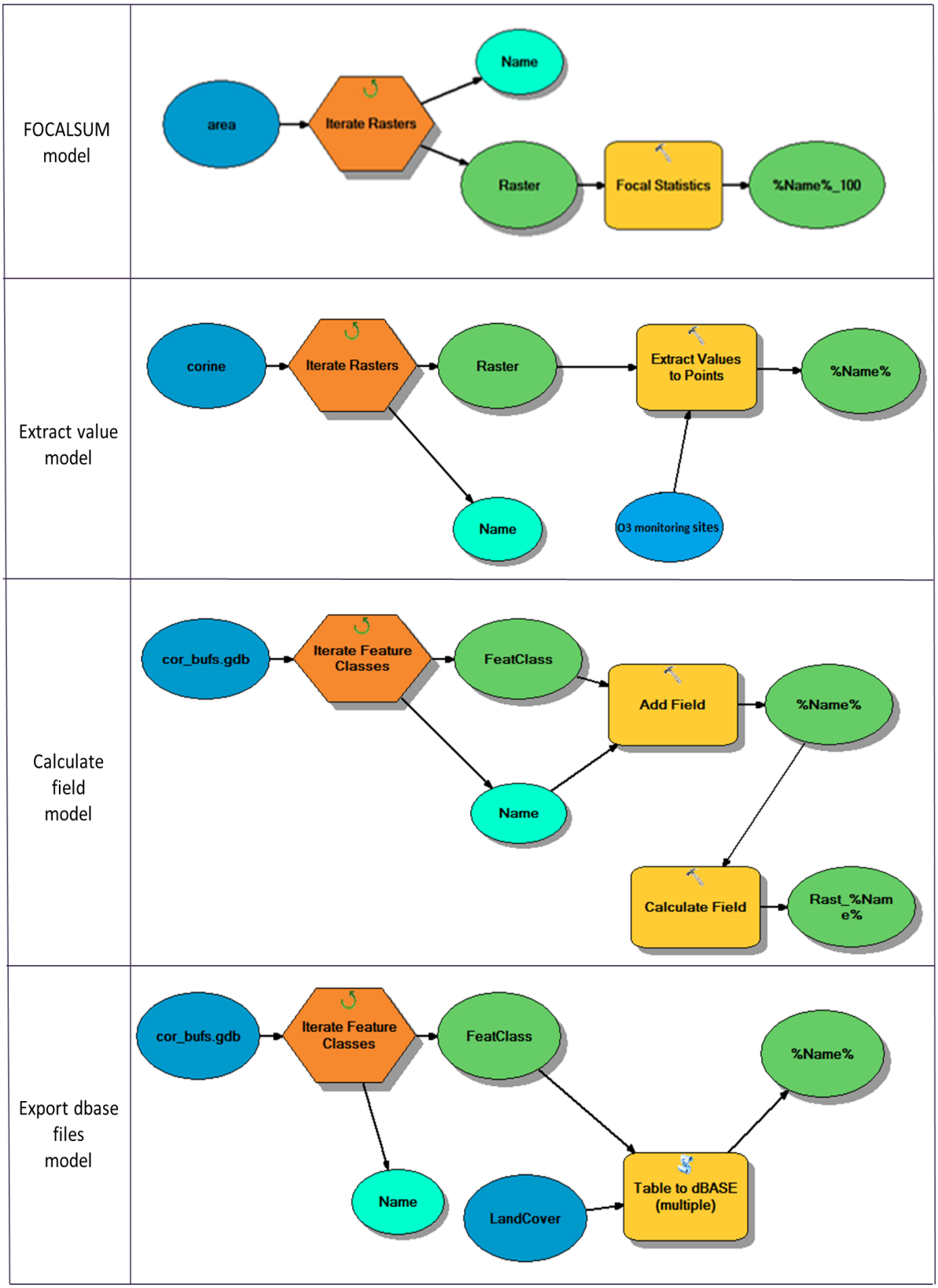


Figure 3.7 Model builder to obtain the different land cover data within different window sizes

The following sections explain the variables that were selected in order to represent and enable modelling of the spatial variations in O_3 across Europe at 100m resolution. The importance of each variable in modelling O_3 concentration is explained, the data source cited, and the original resolution noted. Also, where necessary, any preprocessing (e.g. intersection, interpolation) that was applied to create these 100m GIS-variables is explained in detail.

3.2.3.1 Land cover variables

O_3 is a secondary air pollutant, formed by a series of complex chemical reactions, as outlined in section 2.1.1. Formation and loss are driven by two critical precursors: NOx and VOCs, in the presence of solar radiation (*hv*).

In the absence of detailed emissions data, land cover data were thus included in the analysis as proxies for emissions to the atmosphere, since they describe differences in source type (e.g. industry, residential land, forestry, agriculture) and, to some extent, source intensity (e.g. by defining densely populated areas or heavily trafficked zones).

Land cover nevertheless affects O_3 concentrations in two, opposing ways. Some land cover classes are indicators of O_3 production, because they represent emission sources for O_3 precursors, or situations where favourable conditions for chemical generation of O_3 in the atmosphere may occur. Other land cover types are likely to be associated with reduced O_3 concentrations, because they are related to the release of O_3 scavengers, or encourage deposition of O_3. In practice, these relationships with land cover are often complex and contradictory. In the case of forestry and agricultural land, for example, both these roles may be at work. On the one hand, vegetation acts as a surface for dry deposition, especially during the day when stomata are open (Nowak et al., 2006, Fowler et al., 1998, Massman and Grantz 1995). At night, also, O_3 concentrations tend to decline due to deposition on the soil surface with no substitution by photochemical production. On the other hand, forestry and agriculture can be important emission sources for biogenic VOCs (isoprene and monotorpene), which are more reactive by 2-3 times than anthropogenic VOC (Carter, 1991). This leads to increased O_3 concentrations (Chameides et al., 1988).

Land cover data in 100m resolution were derived from the CORINE Land Cover Map 2000. The database has been created by semi-automatic interpretation of data collected using

satellite-borne sensors and has a spatial resolution of approximately 25 ha. CORINE land cover data (CLC2000) were downloaded from the EEA web site[13]. CLC consists of 44 primary classes (Appendix, A section III). These were combined into 7 more general groups, as demonstrated in Table 3.4. This was done by reclassifying the original CLC grid using the CON (i.e. conditional) function in ArcMap Spatial Analyst to produce a new raster for each of the 7 classes.

These seven classes were specified on the basis of their influence on O_3 formation and dispersion. High density residential land represents areas of high population density, typically associated with more intense anthropogenic emissions of NOx, which scavenges O_3. Low density residential land comprises areas with lower population densities, which are typically associated with lower NOx emissions. Industrial and commercial land includes a range of different areas (e.g. industrial, commercial and construction). In general these can be expected to be sources of emissions of O_3 precursors, such as NOx, CO and anthropogenic VOC, which will either scavenge O_3 or increase formation of O_3.

Table 3.4 Definition of the 7 land cover domains derived as a combination of the 44 CLC classes

Abv.	Land cover variables	Description	CLC Classes[a]
Highdr	High density residential land	Continuous urban fabric	1
Lowdr	Low density residential land	Discontinuous urban fabric	2
Ind/Com	Industrial/commercial land	Industrial, commercial and construction units	4-9
Herb	Herbaceous land	Pastures, natural scrub and herbaceous vegetation	10-11,18,26-29
Agri	Agriculture land	Arable land, crops and heterogeneous agriculture	12-17,19-22
Forst	Forest land	Forest area	23-25
Opsp	Open Space	Beaches, rocks and open space with no vegetation	30-34

a. CLC classes 35 to 44 representing wetland and water bodies were excluded because they do not characterise land

Three types of green area, varying in the density and height of the vegetation, have been defined: herbaceous land comprises areas of very low vegetation, mainly in the form of shrubs or grass, with few trees. Agriculture includes low, permanent or rotating crops

[13] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2000-raster/clc-2000-v13-100m,accessedon March/2010

(excluding grass). Forests consist primarily of dense vegetation with tall and continuous tree cover (broad-leaved forest, coniferous forest, mixed forest). The three classes of green area (herbaceous, agriculture and forest) are the main sources of biogenic VOC, which contributes to O_3 formation, but may also provide active surfaces for dry deposition. Finally, open space comprises areas with no vegetation, such as beaches and rocks.

The FOCALSUM statistic in ArcMap was then applied to the seven land cover classes to produce grid cells for the different window sizes.

3.2.3.2 Topography

Topographic characteristics of the land are important because of their relationship with meteorological factors that might affect the distribution and transportation of O_3. For instance, O_3 concentrations in Europe tend to increase in mountainous areas (Jonson et al., 2006). Topographic exposure, such as the openness or lack thereof, may likewise influence atmospheric temperatures, and hence photochemical reactions, as well as exposure to prevailing winds which may act to accelerate the dispersal, mixing and deposition of O_3. Topography can therefore be used as a proxy for meteorological factors affecting O_3 concentrations.

Three different topographic variables were derived for use in this study: altitude (height above mean sea level), topex (an index of topographic exposure) and distance to sea. Each of these is explained in turn.

1) Altitude

Altitude was obtained from the Shuttle Radar Topographic Mission (SRTM) v4.1 produced by NASA[14], as ASCII files. The SRTM digital elevation data (DEM) is a high quality elevation data set covering over 80% of the globe, and the whole of this study area. It has a horizontal resolution of 90m at the equator, and data are provided in 5° x 5° tiles, in a geographic coordinate system (WGS84 datum). The vertical error of the DEM is stated to be less than 16m. Areas where water or heavy shadow prevented the quantification of elevation are indicated as "no-data".

[14] available at http://www.cgiar-csi.org/data/elevation/item/45-srtm-90m-digital-elevation-database-v41 last accessed: 20 April 2012 in

The ASCII files were converted to integer rasters using the ArcGIS command ASCII TO RASTER. The tiles were then joined into a single raster dataset by using the command MOSIAC. Finally, the raster was re-projected into the chosen based project LAEA. Bilinear interpolation, which is the appropriate algorithm for continuous data, was used with the registration point set as (N400000, E200000). This bilinear interpolation will cause some smoothing of data because the output grid value is the weighted average of the nearest four cells, as shown in Figure 3.8.

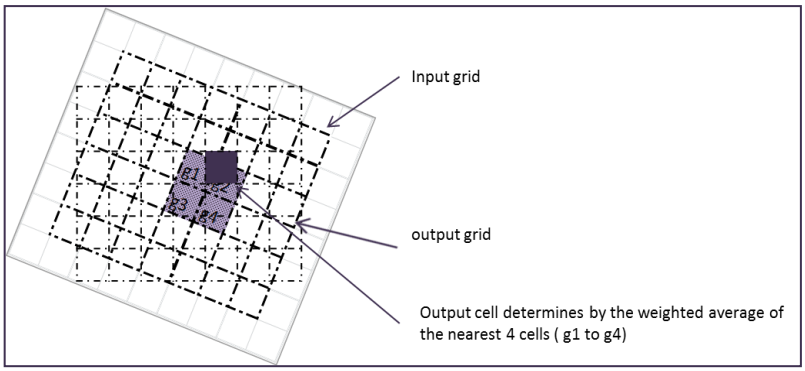


Figure 3.8 Spatial resampling using bilinear interpolation

2) Topex

Topex refers to topographic exposure. It is computed by subtracting the mean altitude of the surrounding area from the altitude at the target area (sum of altitude values within a circular window around the focal cell. The resulting 100m altitude grid (Alt_{100}) was used for this purpose. A distance of 2000m was selected to represent the surrounding area, and was calculated using FOCALSUM (Alt_{2000}). High (positive) Topex values indicate that the target area represents a peak in the landscape, and is therefore relatively exposed; negative values indicate that it occupies a depression or valley, and is therefore sheltered (Figure 3.9). Topex will be zero if the topography is flat.

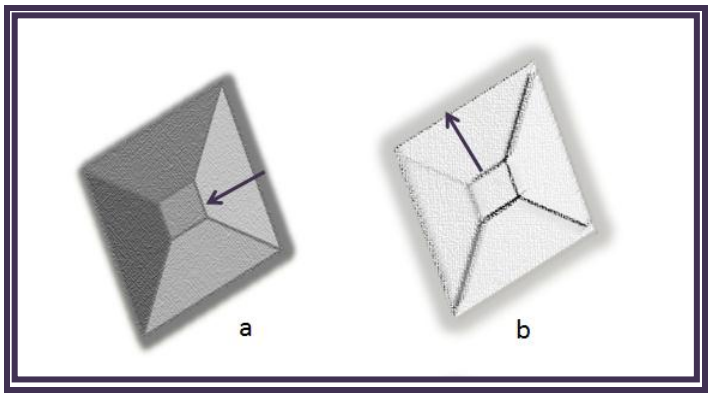


Figure 3.9 Illustrating the positive topex (a) and the negative topex (b)
The direction of the arrow points toward the higher ground between the 100m window (inner-ring) and surrounding neighbourhood (outer-2000m ring)

Topex was specifically calculated as follows:

- Alt_{100} = Altitude for the surrounding 100m window size, based on the altitude grid
- Alt_{2000} = sum of altitude values within a circular window around the focal cell. This window has a radius of 20 cells (a total of 1,257 cells are contained within this window)
- $Alt_{2000\text{-}100}$ = The sum of altitude values in the "outer ring", computed by subtracting altitude at 100m window from the previous result (e.g. Alt_{2000} - ALT_{100})
- $MAlt_{outer}$ = The mean altitude in the "outer ring", computed by dividing the value of the "outer ring" by number of cells (eg. $Alt_{2000\text{-}100}$ /1257)
- Topex = The difference in altitude between the 100m window and surrounding area (e.g. Alt_{100} – $MAlt_{outer}$)

3.2.3.3 Distance to sea

Coastal areas comprise a distinctive O_3 environment and are an important source of O_3 precursors as these are areas of high photochemical activity. In addition to often being densely populated areas, there are several factors which act to influence O_3 concentrations, including short and long range transport, local emissions, meteorological phenomena influencing transport, dispersion and recirculation of pollutants, and photochemical activity. Recycling and trapping of pollution originates from the generally large heat differences

between ocean and land which produce a movement known as sea breeze circulation. Both horizontal and vertical coastal recirculation can occur, which can affect the air quality. Horizontal recirculation returns the air mass to its source area the next day, whereas the vertical recirculation currents return the air a few hundred metres down onto the land surface (Hsu, 1988). These factors increase O_3 concentrations in coastal areas, by bringing in sea breezes enriched with O_3 (Klingberg et al., 2012). Distance to sea, therefore, provides a potential proxy for these effects. Distance is computed as the straight line distance to the nearest body of open sea.

The following steps were used to create this data set (see appendix A, Section IV for full details).

1. The raster coastline from CORINE2000 (class 523) was converted into a coverage using the command CONVERSION. This was buffered by 20km to represent the boundary to the open water.

2. It is computationally intensive to compute the distance from each 100m cell to the coast; therefore centroids for a 1km grid for Europe were used instead. These were stored as coverage.

3. The NEAR command was used to compute the distance (in metres) from each 1km centroid to the nearest open water.

4. Distance to ocean, based on the 1km centroids, was then interpolated to the 100m level using inverse distance weighting and stored as a raster. Values from the resulting 100m distance to sea raster were extracted for the O_3 monitoring sites using Extract values to points.

5. This method for interpolating distance to sea was validated at the monitoring sites, by directly calculating the distance using the command NEAR between sites and open water. The correlation was found to be 0.99 at the monitoring sites.

3.2.3.4 Road length

Because a large proportion of the NOx emitted in Europe derives from road transport, data on road length for different road classes (local, secondary and major roads) were also obtained, to give a proxy for scavenging by transport-related NOx.

Road data were obtained from Eurostreets version 3.1, which is a 1:10,000 digital road network based on the TeleAtlas MultiNet TM. These data were obtained through the European Study of Cohorts for Air Pollution Effects (ESCAPE) project, and were converted from vector to a 100m raster by colleagues at Imperial College[15]. Eurostreets does not include traffic intensity data; however it does include a road classification (FRC code). To simplify the classification, FRC was reclassified into three groups as illustrated in Table 3.5.

Table 3.5 Selected road classes based on Eurostreets road classes

FRC	Road classes	New classes	Abbreviation
0	Motorways	Major roads	MR
1	Roads not belonging to main road Major importance		
2	Other Major roads		
3	Seconds roads	Secondary roads	SR
4	Local connecting roads	Local roads	LR
5	Local roads of high importance		
6	Local roads		

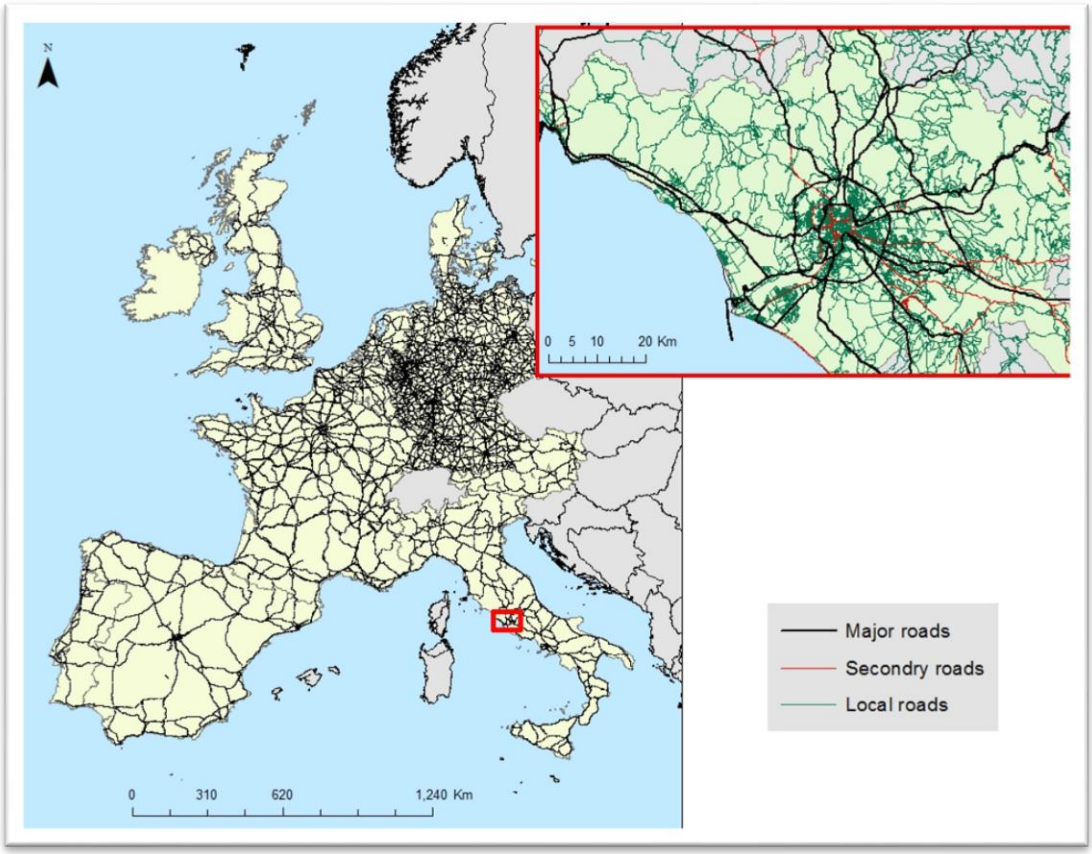


Figure 3.10 Road classes
Derived from Eurostreets version 3.1

[15] Vienneau and Lee.

Figure 3.10 shows the map of the whole study area, depicting only the first class, major road (MR); the inset for the city of Rome shows all three road types.

On obtaining the rasters for each road class (FRC), the following additional steps were performed to prepare these data for use in this study: firstly, reproject roads into the LAEA projection and create the 100m base polygon (grid) shapefile; secondly intersect road vectors with the base polygon; thirdly sum road length by FRC for each 100m polygon area; finally convert the polygons to rasters. These were then combined to create three road classes, as shown in Table 3.5. This was done using the PLUS command in ArcGIS. As per other variables, the FOCALSUM statistic was then applied to the three resulting road grids to produce grids for the different window sizes.

3.2.3.5 Meteorological factors

Important surface meteorological factors related to O_3 concentration are cited as temperature, wind speed, solar radiation, and precipitation (Tarasova and Karpetchko, 2003, Dueñas et al., 2002, Lou Thompson et al., 2001, Rodríguez and Guerra, 2001, Dabdub et al., 1999). All four were used in this study.

The role of meteorological factors in O_3 formation and dispersion can be summarised as follows. In general, high O_3 concentrations are observed in favourable photochemical conditions, characterised by high temperature, high solar radiation (i.e. sunny) and in the presence of O_3 precursors. In contrast, in overcast or rainy conditions, characterised by high total precipitation and low sunlight due to the cloudiness, as well as low temperatures, O_3 concentration is low due to the slow rate of photochemical reactions and to loss of O_3 by wet deposition (Andersson et al., 2007, Lelieveld and Crutzen, 1991).

The effect of wind speed on O_3 concentrations is more complex, and depends on the specific atmospheric conditions. One effect is to reduce O_3 concentrations by encouraging dispersion away from O_3-rich areas. Downwind of these sources, however, the wind has the reverse effect, of bringing in more O_3-enriched air. In the vertical dimension, equally, contrasting effects may occur. Where the boundary (i.e. ground) layer acts as a source of O_3 due to chemical generation, increasing wind speed reduces surface-level O_3 concentration by generating turbulence and increasing vertical mixing. Conversely, if the O_3 chemical budget

in the boundary layer is negative (i.e. if O_3 concentrations are greater at higher levels in the atmosphere), vertical transport transfers O_3-rich air from aloft downward (Elampari and Chithambarathanu, 2011, Shan et al., 2009, Tarasova and Karpetchko, 2003, Davies et al., 1992). In addition, scavenging by NO may be reduced under conditions of high wind speed (Dueñas et al., 2002, Rodríguez and Guerra, 2001, Dabdub et al., 1999). Surface O_3 concentrations may therefore show either a negative or positive correlate correlation the wind speed depending on the distribution of O_3 production and of scavenging, both horizontally and vertically in the atmosphere, relative to the measurement site.

Global meteorological data at a spatial resolution of ca. 40 km were downloaded from European Centre for Medium-Range Weather Forecast (ECMWF) website[16] as a Network Common Data Form (NETCDF) files[17]. These data are produced by the latest ECMWF global atmospheric analysis (ERA-Interim) for the period of 1989 to present. The data required for this study derived from two main sources: analyses data contain the monthly means from March 2001 to Feb 2007 based on daily means, for wind speed (ws) in m/s and temperature (tmp) in °C, while the forecast data contains surface solar radiation (ssr) in W/s and total precipitation (tp) in mm.

In ArcMap, NETCDF files can be opened as a layer using multidimension tools in ArcTool Box. These can then be easily exported as dBase files for further processing. In SPSS, the monthly data for each variable were used to calculate the annual, summer and winter means for each year. Summer covered the months from June to August, while winter covered December to February. These were saved as a dBase file to be opened in ArcMap.

As the global meteorological data were at a 40km resolution, a simple interpolation, using the square of the inverse distance, was employed to smooth the data to 100m. First, however, the 40km centroids were reprojected into LAEA. Figure 3.11 demonstrates the processing in ArcMap.

Global meteorological data were needed for the spatial model as this predicts the long-term mean of O_3 cross western Europe. At a later stage in the research, a space-time model is applied at a country or city level. For this, more local meteorological data were acquired to

[16] http://www.ecmwf.int/services/archive/d/edit/personal/temporary/netcdf, accessed on 2/2011

[17] Type of data format widely used in the atmospheric science and oceanography to sort multidimentioned arrays of data

reflect the daily variation (non-systematic variation) which obtained from meteorological stations in the country or city of interest.

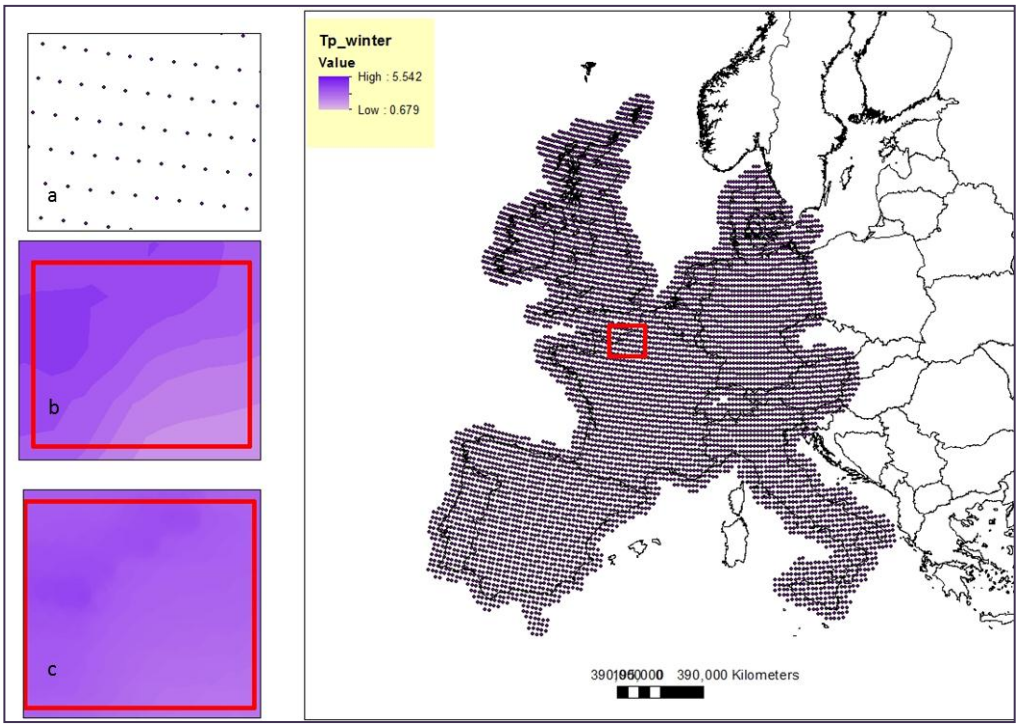


Figure 3.11 Illustration of the steps to convert meteorological data from 40km (point) to 100m grid (raster) a) the points with 40 km resolution, b) smooth met data using IDW interpolation method, and c) save result as a 100m raster

3.3 Summary of available data

As described in the previous sections, all the variables needed to develop the spatial model were processed and stored as 100m regular grids for the whole study area. The final database included all the variables shown in Table 3.2, within windows with a radius of 100m, 300m, 500m, 1km, 5km, and 10km. The final database also includes a cleaned hourly O_3 concentration at the 1,211 sites meeting quality criteria and maintained in this study.

It might be noted that converting all these data into a common 100m raster GIS required considerable effort, due to the heavy computational load involved. The effort, however, is repaid in the subsequent analysis, which is greatly accelerated by the availability of consistent, high resolution, ready-made data sets.

4 O_3 sites classification

The aim of this chapter is to classify O_3 monitoring sites in west Europe based on their temporal pollution signature, as a precursor to temporal modelling of concentrations at O_3 monitoring sites. The relationship between the resulting site types and selected environmental variables (e.g. land cover, topography, and climate) is then explored as a basis for predicting site type membership of unmonitored locations (i.e. across a 100 metre grid).

4.1 Introduction

Tropospheric O_3 is a complex pollutant influenced not only by emissions, but also by long range transport of O_3, surface deposition, and photochemical activity. The complex nature of O_3 means that the classification of monitoring locations into groups, exhibiting similar temporal patterns, is not a straightforward process.

This is an important element of the research because data are available for a large number of O_3 monitoring sites (1211), distributed somewhat unevenly across west European countries. A global model of temporal variations in O_3 concentrations, fitted to all these sites, is unlikely to be equally effective across this study area (e.g. both in rural areas with no local emission sources, and in areas close to busy main roads, where pollution is affected by the time-varying flow of traffic). On the other hand, developing a discrete model for each station is computationally uneconomic, and in any case does not greatly help in estimating concentrations at unmonitored sites (in this case for a 100m grid), since these cannot easily be attributed to a specific monitoring site (and therefore a specific time model). Instead, it is clear that modelling is best done for groups of sites, classified according to their specific temporal O_3 signatures.

One possible classification is that assigned to the monitoring sites within the European network, reported in the AIRBASE database. Sites within AIRBASE are described on the basis of two classifications: one referred to as site type (background, traffic and industrial) and the other, site location (urban, suburban, rural and unknown), as shown in Table 4.1. Background stations are also divided into subclasses: near city, regional and remote background stations. Together, these classifications are intended to define the

characteristics of the areas around each site, and which the site is meant to represent. It should also be noted that some sites in the AIRBASE dataset are classified as unknown, where relevant metadata have not been submitted, making classification impossible.

The representivity of monitoring sites, and the extent of the areas that they actually represent, is difficult to determine. In principle, pollutant concentrations in the area surrounding any given site should not differ by more than a small and specified amount from the values measured at that site. The radius of the area within which this is likely to be true, however, may range from a few metres for sites at traffic hot-spots through to tens of kilometres for regional background sites. As explained in the report describing the criteria for selecting sites for AIRBASE, however, this is rarely assessed (Larssen et al., 1999).

Table 4.1 Characteristics for classifying monitoring stations for AIRBASE (Garber et al., 2002)

Type of zone	
urban	Station is located in a city (i.e. continuously built-up area)
suburban	Largely built-up area (i.e. continuous settlement of detached building mixed with non-urbanized area (e.g. agriculture, lake))
rural	Area does not fulfil the urban or suburban criteria.
unknown	Metadata have not been submitted
Type of station based on dominant emission	
traffic	Located near traffic sources
industrial	Located near single industrial sources or industrial areas with many sources
background	Level is not determined significantly by any single source or street, but by the integrated contribution from all sources upwind of the station.
unknown	Metadata have not been submitted

To avoid subjectivity, and to be readily reproducible, site classification should ideally be based on relevant and available, largely quantitative, information. The classification should also be able to distinguish between areas which differ in terms of their pollution levels. In practice, creating a quantitative and reproducible classification is not possible, due to the highly localised nature of variations in pollutant concentrations, especially in urban areas or areas close to the sources of O_3 precursors. Delimiting representative zones thus requires intensive measurements around the monitoring station, and by the very nature of the monitoring networks, these do not exist. Another problem is that different countries or agencies may interpret the criteria for classification in slightly different ways, so that the classification is not applied entirely consistently. In the case of AIRBASE, the classification

was also derived regardless of the type of pollutant. Different pollutants, however, are likely to show different patterns of variation, and different relationships with emission sources. In practice, therefore, different criteria may be needed to classify sites for different pollutants.

For the aims of this research – i.e. to model space-time variations in air pollution – a further limitation exists in the AIRBASE classification. This is that the classification is essentially aimed at distinguishing only geographic variations in air pollution, rather than variations over time. There is no guarantee, therefore, that the sites in any one category will display similar temporal pollutant patterns. This is only likely to be achieved if classification is explicitly based on, or calibrated against, time-varying profiles of pollution measured at a representative sample of sites.

To illustrate this, the monitored hourly data from March 2001 to February 2002 from 1253 sites were analysed as a pilot analysis, to explore the spatial variations in observed concentrations using AIRBASE site type classification. This was done using variance components analysis (VCA). VCA is a way to assess the quantity of variation in a dependent variable (in this case, ambient O_3 concentration) in relation to one or more affects variables. The fundamental output is a variance components table which shows the percentage of variance attributable to the main effect of the variables.

Spatial variables were categorised on a hierarchical basis in this analysis. Site type is based on a simple binary classification. At the highest level (site type A) all sites were classified as either background (i.e. in rural or urban background areas, unaffected by local emission sources) or other. At the next level (site type B), the 'other' category was further subdivided. Finally, Site represents the unique effects of each monitoring site. This was possible only in three countries (Belgium, Portugal and Spain) which have a reasonable number of sites (more than 10% of the total number for each site type) in each category of site type B (i.e. background, industrial, traffic and unknown). Results are summarized in Table 4.2.

Table 4.2 Percentages of variation explained by spatial factors.

countries	Site type A	Site type B	Site	Total
Spain	29	9	62	100
Portugal	23	14	63	100
Belgium	41	3	56	100

As Table 4.2 shows, Site type A accounts for 23-41% of the spatial variation, suggesting that these two categories of site are distinctly different. Differences in the 'other' category (e.g. between industrial and traffic sites) have only a small effect, explaining 3-14% of the spatial variability. Much of the remaining variation (ca. 60%) was attributable to Site: i.e. represented temporal variability within individual sites. These unique differences between sites clearly cannot be explained by the AIRBASE site type. All these factors limit the utility of the AIRBASE classification as a basis for modelling short term variability.

A number of previous studies (Joly and Peuch, 2012, Flemming et al., 2005, Snel, 2004, McGregor, 1996, Mcgregor and Bamzelis, 1995) have pursued the objective of classifying pollutant-specific monitoring sites on the basis of the temporal variations in measured concentrations of the monitored pollutant. Mcgregor and Bamzelis (1995), for example applied principal component analysis (PCA) and cluster analysis to derive airmass types based on meteorological data. The original dataset consisted of eleven different meteorological factors (e.g. wind speed and direction, different elements of temperature, cloud cover, solar radiation) for 365 days. Daily averages were used in PCA in seeking a combination of meteorological factors that grouped together at a location for a given time. This produced four components, defined as westerly flow, cloud, fog and hygrothermal conditions. These components were then subjected to hierarchal cluster analysis (HCA) to identify groups of days with similar meteorological conditions. Six site types were identified in terms of the air pollution and synoptic situation they represent. In a later study, McGregor (1996) used the same approach to determine whether similar air quality patterns existed in the urban area of Birmingham. Data from seventeen SO_2 monitoring sites were used, comprising 571 daily SO_2 measurements in the winter of 1979-80. The PCA reduced these data to four components representing the temporal behaviour of SO_2. The monitoring sites were then grouped by HCA based on these components. As a result, four different site types were identified, ranging from heavily polluted to low pollution areas.

Snel (2004), in a pilot study, used weekly NO:NO_2 ratios for summer and winter seasons for three years based on hourly data for 1999, 2000 and 2001, to group 465 O_3 monitoring sites across Europe. The NO:NO_2 ratios (6 ratios: 2 seasons x 3 years) were then used in K-mean cluster analysis to classify all AIRBASE sites into three site types: rural background, urban background and traffic, using the available NO and NO_2 data. In many of the twenty three

countries analysed, more than half of sites classified in this way differed from the registered site type in AIRBASE.

Flemming et al. (2005) classified 650 air quality sites in Germany for four critical air pollutants (O_3, NO_2, SO_2, and PM_{10}) on the basis of hourly time series data from 1995 to 2005 (ca. 2-3000 data points). The procedure for classifying the O_3 monitoring sites was, first, to scale the time series data using log-medians of the daily average (P50DA) and normalized daily variability (P50DV). This was done despite the fact that they were negatively correlated. HCA was then run, producing six different O_3 site types. The stability of these six site types was tested by cross validation, comparing the results with a reference classification produced in another study. Flemming et al. (2005) claimed an excellent agreement but do not report the actual percentages.

In the most recent study, conducted by Joly and Peuch (2012), 4956 air quality sites were classified across 35 European countries based on hourly time series data over the period 2002-2009 for different air pollutants (O_3, NO_2, NO, SO_2, and PM_{10}). Time series data for each pollutant were described by eight metrics:

- daily maximum;
- daily amplitude (daily maximum minus daily minimum) for each month of the year;
- the annual mean;
- the summer minus winter mean;
- the high frequency standard deviation;
- weekend effect on:
    - the daily mean,
    - daily maximum,
    - Standard deviation.

Classification was constrained to be consistent with the AIRBASE meta-data. Therefore, to distinguish between rural and polluted sites (i.e. traffic and urban), discriminant analysis (DA) was employed using the eight metrics to classify monitoring sites to rural and urban based on AIRBASE meta-data. The DA results were then arbitrarily stratified using the nine percentiles (10 to 90%) as fixed thresholds to produce ten site types. Comparing the ten site types with AIRBASE site types showed considerable inconsistency between the two classifications. The authors also stated that classifying the air quality sites within Europe is not straightforward as the measurements represent different environments.

All these studies agree with the notion that classification of sites within any given network should be pollutant specific (e.g. classification of sites based on PM may not be the best classification with respect to O_3). They also agree that this classification should be based on a thorough understanding of measured concentrations over a period of time. On the other hand, it is evident from these studies that the AIRBASE does not meet these criteria, due to the subjective method followed in the classification, regardless of the type of pollutant. This emphasises the need for a more objective approach to classify the O_3 sites.

The main difference between the studies described above was in the choice of the parameters used to condense the information from the time series data into a form suitable for classification. This was done either through the use of indicators (e.g. NO:NO_2 ratios, threshold NO:NO_2 ratios, P50DA and P50DV) or by using statistical methods to reduce the massive time series data to a number of components which explain most of the information in the data (i.e. components derived from PCA). The studies agree again, however, in using cluster analysis to classify the monitoring sites (i.e. to group sites with similar temporal behaviour) based on these parameters.

A further question remains at the end of these studies – namely, how to assign the site type to unmonitored location over the study area. As Flemming et al. (2005) have noted, one way of doing this is by cluster analysis, using the 'centroid approach', which employs the conceptual distance between the group centres as a cluster criterion (Wilks, 1955). This is only possible, however, where the same attributes and data used to set up the initial classification (in this case O_3 concentration) is also available for the target locations. In this instance, this is clearly not the case, since the target locations are unmonitored, and the only available measurements are the environmental data.

This is a common situation when mapping or modelling environmental phenomena. It is encountered, for example, in soil science (i.e. digital soil mapping), where the aim is often to map soil properties, using data that are available for only a subset of sample locations. Typically it is done by establishing a relationship between the properties of interest and the soil class, then using data on soil class to estimate the relevant soil properties at any location (Abdel-Kader, 2011, Debella-Gilo and Etzelmüller, 2009, Kempen et al., 2009).

One context in which this arises is updating national soil maps. Field surveys are very costly, so some form of predictive method is needed which allows maps to be updated on the basis

of a relatively small amount of sample data. For this purpose, Kempen et al. (2009), for example, aimed to assign soils to ten major soil classes at unvisited locations across Drenthe in the Netherlands. Using multinomial logistic regression (MLOR), the relationship between twelve environmental datasets – elevation, ground water classes, historic and recent land cover classes, geomorphologic classes, and paleogeography classes - and the soil classes was quantified, from surveyed sites. The environmental data were selected based on expert knowledge and ten models were built. From these models, the probability of occurrence of the ten site classes was estimated and used to assign the unvisited sites, across a 25m grid. Results were assessed by estimating the purity of the national soil map (i.e. the percentage of sites in each soil map unit that conformed to the designated class as assessed by the validation sample): based on this measure 58% of sites were found to be correctly classified.

Following the same approach, Abdel-Kader (2011) used soil map classes defined by field survey to predict the distribution of soil classes in the north coast of Egypt. A set of parameters (i.e. information about the soil and terrain) were calculated and then entered into MLOR to derive probabilities of soil class membership. The probability models thus developed were used to predict the spatial distribution of the soil mapping units at grid resolutions of 28.5 m × 28.5 m and 90 m × 90 m in adjacent, unvisited areas at Matrouh and Alamin.

Following the same concept, Beelen et al. (2009) classified urban and rural areas across a 1km grid for Europe. Classification was done by intersecting the monitoring sites with CORINE land cover data, and using DA to generate a discriminant function, distinguishing between urban and rural sites. Two environmental variables were developed for each of two different window sizes (1km* 1km and 3km*3km) on the basis of land cover data. The area of total built-up land was used to characterise urban areas in each of these windows, and the sum of agricultural and forest classes to characterise the rural area. Using these four variables, the overall accuracy of classification was 86% for the urban area and 89% for the rural area, compared to the AIRBASE classification.

These examples of studies suggest that MLOR and DA are appropriate methods to generate discriminant function(s) defining the site types in terms of their environmental factors. Both methods assume a linear relationship between continuous predictors (selected environmental data) and the categorical variable (different site types).

4.2 Methodology and results

In this study, classification of the O_3 monitoring sites was done by using the hourly time series concentrations from the 1211 sites in the AIRBASE network which met pre-specified data quality/capture criteria (section 3.2.2), together with the environmental data developed in Section (3.2.3). As illustrated in Figure 4.1, analysis comprised four key stages:

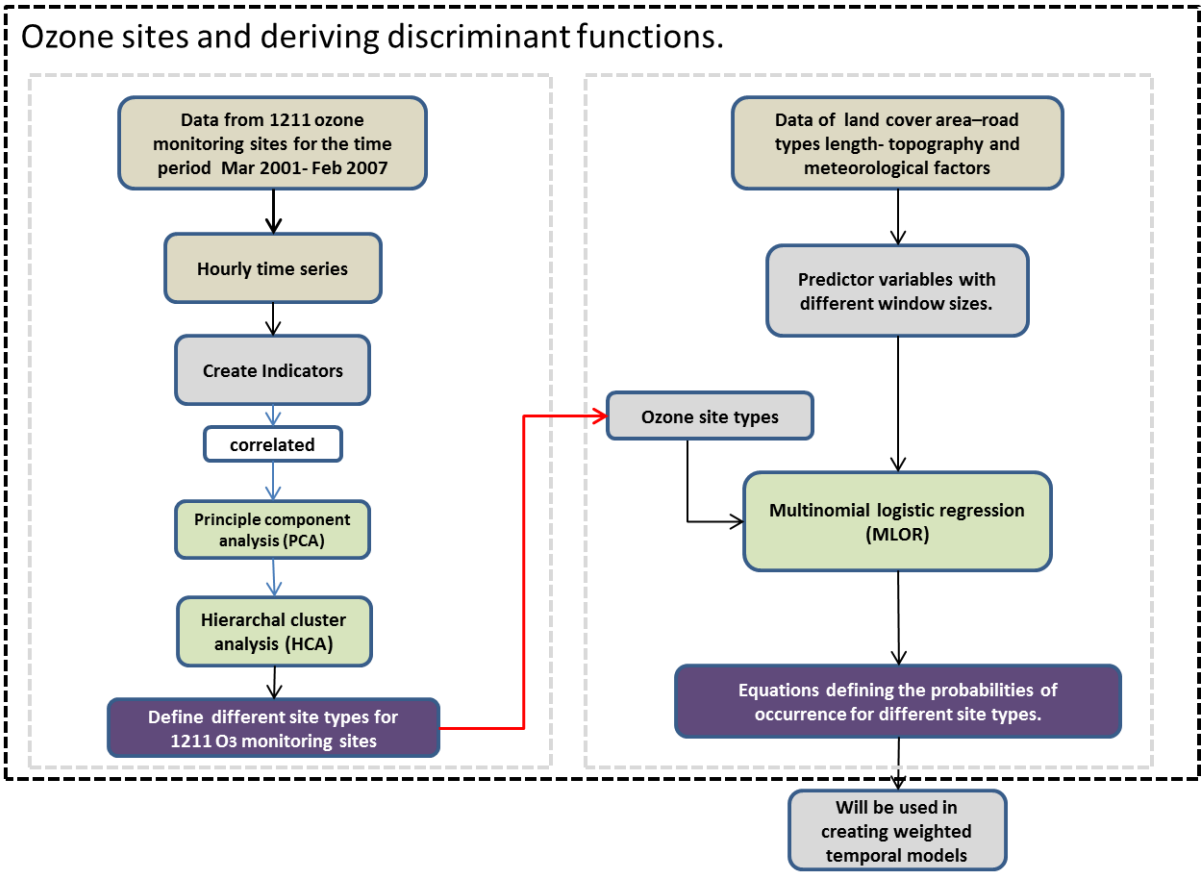


Figure 4.1 The procedure for classifying the 1211 monitoring sites and deriving discriminant functions

1. Creation of the indicators (Section 4.2.1). This involved defining and selecting suitable indicators (parameters) which potentially have the ability to discriminate between O_3 monitoring sites in terms of their temporal patterns, over different time periods (diurnal, weekday/weekend and seasonally).
2. PCA (Section 4.2.2). The indicators created in step 1 were correlated, so PCA was used to produce uncorrelated principal components, explaining most of the information in the indicators.
3. Classification (Section 4.2.3). HCA was then applied to the principal components to classify the 1211 O_3 monitoring sites into a smaller number of groups with distinct temporal patterns.
4. Definition of site types in terms of environmental factors (section 4.2.4). This was done by establishing the relationship between selected environmental variables and site types using discriminant functions in MLOR. These functions will later be used to predict the probabilities of site-type membership for unmonitored locations.

The site type classification thus generated will enable the development of different temporal models for different sites, depending on their probability of site type membership.

4.2.1 Creation of indicators

As noted above, classification of monitoring sites should be based on a thorough understanding of measured concentrations over a period of time. This purpose can be achieved by creating indicators which characterise the temporal variations in the measured concentrations.

The purpose of the indicators in this study, simply, is to provide a means of describing the temporal 'signatures' (i.e. repeated patterns) of O_3 concentrations that reflect different types of O_3 behaviour. For this purpose, the indicators should describe the hourly record (or trace) of concentrations at any site in a way that allows it to be compared with that at another site. The indicators should thus facilitate grouping of monitoring sites in a way that allows specific time models to be derived for each group of sites. Ideally, also, it should be possible to identify and map these groups on the basis of relevant exogenous, environmental variables.

The indicators must, therefore, be selected to represent the most important systematic time-patterns (variations) in the O_3 data. These variations include, especially:

1. Seasonal variations (winter-summer contrasts), due mainly to broad-scale climatic and topographic effects (e.g. degree of continentality, or mountain effects);
2. Weekday/weekend variations, due, for example, to the weekly cycle of emissions associated with industrial activity and traffic;
3. Diurnal variations (morning and evening peaks, and night-time levels) probably reflecting the effects of local traffic (especially of NOx), photochemical processes, and inter-regional transport of O_3.

Because these temporal variations are themselves due, in part, to the effect of the surrounding environment, identifying signatures in this way should reflect differences in site type: for example between transport and non-transport; urban, suburban and rural; windy and calm; or warm and cool environments - and all the different combinations of these that might be expected to occur in the real world. By the same token, they should also relate to measurable environmental characteristics (e.g. topography, meteorology, land cover) that can then be used to assign unmonitored locations to the same site types.

The variations in O_3 concentrations to be captured by these indicators can be characterized in terms of different statistical metrics. The most useful metrics are likely to be measures of central tendency both overall and periodically (i.e. within specified averaging periods, such as seasons, weeks, days), periodic extremes (to show, for example, whether the pollution peaks within these periods are especially marked), and periodic variability (to demonstrate whether levels remain broadly constant or fluctuate within the specified periods). For O_3 three key periods (or time scales) are especially relevant: seasonal (summer (SUM) versus winter (WINT)), hebdomonal (weekday (WD) versus weekend (WE)), and diurnal (morning (AM), afternoon (PM), and night (NI)).

The methodology used was devised on the basis of a further analysis using VCA analysis, as mentioned in Section 4.1. This time a full range of spatial and temporal factors were considered to assess the quantity of variation in those factors. Spatial factors were represented by site type A (rural and urban background versus other) and Site; as there are not enough traffic and industrial sites in all countries, site type B was excluded in this analysis.

Temporal factors were generated by aggregating the hourly concentrations from the 1253 sites for the year from March 2001 to February 2002 to the following time intervals: hour of the day[18], time of day (afternoon versus rest of the day)[19], day of the week[20], weekday/weekend (weekend versus weekdays)[21] and season[22]. These were all defined as fixed effect factors for a random sample of monitoring sites; site was therefore regarded as a random effect factor. Results are summarized in Table 4.3. These effects variables, it should be noted, are hierarchical, in that sites nest within site types, and hours within times of day, within day of week, within weekday/weekend, within season.

Table 4.3 Percentage of total variance in ambient O_3 concentration attributable to spatial and temporal factors

Countries	Site type A	Site	Total spatial	Season	Weekday/ weekend	Day of week	Time of day	Hour of day	Total temporal	Error	Total
Ireland	8.9	9.6	18.5	15.3	0.0	0.1	1.7	0.5	17.6	63.9	100
United Kingdom	0.7	13.9	14.6	11.8	0.1	0.4	3.4	1.3	17.0	68.4	100
Denmark	6.2	3.4	9.6	16.8	0.2	0.4	3.5	1.8	22.7	67.7	100
Portugal	3.9	12.9	16.8	10.1	0.5	0.4	10.6	2.3	23.9	59.2	100
Spain	8.4	20.1	28.5	13.1	0.3	0.0	9.8	2.8	26.0	45.5	100
Austria	1.8	18.3	20.1	23.2	0.1	0.1	6.8	2.2	32.4	47.5	100
Belgium	3.2	4.5	7.7	21.4	0.1	0.1	8.4	2.9	32.9	59.3	100
The Netherlands	3.5	2.3	5.8	20.8	0.1	0.1	8.3	2.1	31.4	62.8	100
France	0.1	9.2	9.3	20.7	0.1	0.1	9.2	3.9	34.0	56.8	100
Germany	1.2	10.0	11.2	22.4	0.2	0.2	9.0	2.3	34.0	54.6	100
Italy	2.0	12.7	14.7	23.0	0.3	0.0	9.6	3.0	35.9	49.3	100
On Average%	3.6	10.6	14.2	18.1	0.2	0.2	7.3	2.3	27.9	57.8	100

[18] Hour of day: O_3 conc. for each hour from 1AM to 12AM (24 hours).
[19] Time of day: 12PM THRU 19PM= 1, ELSE=2.
[20] Day of the week: MON=1, TUE=2, WED=3,THU=4,FRI=5,SAT=6 & SAT=7.
[21] Weekday/weekend: Saturday and Sunday=2, ELSE=1.
[22] Season: winter: Dec to Feb, spring: Mar to May, summer: June to Aug and autumn: Sep to Nov.

Table 4.3 shows that, cross all the eleven countries studied, 14% of the total variability was associated with the spatial factor; 28% with temporal factors, and the remaining 58% of the variability was unpredictable (i.e. error).

The proportion of spatial variation ranges from 6-29%. In general it is weakest in countries that are relatively flat and uniform (the Netherlands, Belgium and Denmark). In these countries, too, site tends to account for a relatively small proportion of the spatial variation (over half in the Netherlands and Denmark). The largest spatial effects occur in Spain and Austria, both topographically diverse countries, and Ireland. The last of these has a relatively small number of sites, split between the larger urban areas (e.g. Dublin) and remote rural areas. In this country, therefore, site type A is again relatively strong.

Temporal factors account for between 17% of the total variation in the O_3 data in Great Britain (GB) and 36% in Italy. Most of the temporal effect is due to season, with a moderate proportion also associated with time of day, and a somewhat smaller proportion with hour of day. This indicates that the temporal patterns comprise both a short and long-term cycle of variation. The difference in the proportion of variance explained by temporal factors is also of note. GB, for example, shows a lower proportion of temporal variation (17%), perhaps reflecting its maritime environment, which experiences little seasonal variation in weather conditions and thus probably has relatively uniform O_3 concentrations over time. In Italy, in contrast, temporal variation accounts for 36% of the observed variability, with a strong seasonal component, characteristic of its Mediterranean climate, with much greater temperature extremes, both between seasons and at a diurnal scale. Variations in O_3 concentrations attributable to day of week, on the other hand, are very small: 0.1% in GB, and 0.3% of the total variance in Italy. This suggests that the weekly cycle of work activity (and associated emissions) is not a major factor.

The large amount of unexplained variation in all the eleven countries (58% on average) is also notable. This indicates that the hourly O_3 data contain a large amount of noise, which will be difficult to model without the use of additional information (e.g. on local emissions and meteorology). Likewise, it might be expected that only about a quarter of the overall variability will be explainable using temporal (i.e. seasonal, weekday/weekend, and diurnal) indicators, as proposed here.

Line graphs were also drawn for a random subsample of sites to explore the diurnal pattern of variation, and to define the timing of the troughs and peaks that occurred (Appendix A, Section V, Figure A.1). This, therefore, helped to determine how best to specify the 'time of day' factors and to see whether this varied systematically between sites.

Results suggested that the afternoon peak typically occurs between ca. 14.00 and 18.00 hours in both GB and Italy. In some sites, another, secondary peak occurs at night between ca. 2.00 and 5.00 hours. Evidence of the effects of scavenging of O_3 by NO is also clear in the graphs, for a clear trough in the concentrations typically occurs between ca. 6.00 and 10.00 hours, coinciding with the morning rush hour. Three critical periods of the day were thus defined: afternoon, from 13.00 to 19.00; night, from 22.00 to 05.00; and morning from 06.00 to 12.00.

On this basis, it was concluded that indicators were needed to reflect the average, extreme and within variability concentrations as follow:

1. **Time of day indicators**, to reflect hour-to-hour variations within a day - and based on hourly averages.
2. **Day of week indicators** to reflect day-to-day variations within a week - and based on daily averages.
3. **Seasonal indicators** to reflect week-to-week variations within a season - and based on weekly averages.

In order to enable meaningful comparisons between sites, it is essential to normalize the indicators; otherwise, the data are likely to be dominated by differences in the overall average concentration (i.e. between polluted and less polluted sites) rather than the temporal variation at a site. Normalization was therefore done against the long-term mean concentration for each site, as illustrated in Table 4.4. This was chosen rather than the standard deviation (SD), because sites which do not vary much overall (have a low annual SD) might appear to vary hugely within some of these time periods in relative terms (because the denominator, the annual SD, will be very small). In practice, however, the mean and SD are usually very closely correlated.

The diurnal (time of day) indicators are based on hourly averages. To minimise high correlations between each of these indicators, the 24 hours of the day were split into three periods, as defined above. These are meant to be reasonably homogeneous in term of the

influences acting on O_3 concentration and representative of some sort of repeated cycle in the daily pattern of variation.

The weekday/weekend indicators are constructed at the next level of aggregation, and are based on daily averages. A week is taken to be from Monday to Sunday. During the working week (Monday to Friday) O_3 concentrations vary little from day to day. Weekday versus weekend concentrations, however, tend to vary substantially (Mayer, 1999). For instance, during the weekday, emissions from traffic and industrial sectors (basically NO) are likely to be higher than at the weekend. In urban sites, especially, concentration variations between weekdays and weekends may be expected to be very clear. On the other hand, at any rural site far from these emissions, this variation is likely to be limited.

Day of week indicators are thus created for two periods: weekday and weekend, partly to help separate areas unaffected by local pollution sources (rural) from those with a marked source effect (traffic and urban sites).

Table 4.4 Indicator labels and formulae

Type	Indicator Label	Formula
Average	SUM_Nmean	(SUM_mean-Weekly_mean)/Weekly_mean
	WINT_Nmean	(WINT_mean-Weekly_mean)/Weekly_mean
	WD_Nmean	(WD_mean-Daily_mean)/Daily_mean
	WE_Nmean	(WE_mean-Daily_mean)/Daily_mean
	AM_Nmean	(AM_mean-HOURLY_mean)/HOURLY_mean
	PM_Nmean	(PM_mean-HOURLY_mean)/HOURLY_mean
	NI_Nmean	(NI_mean-HOURLY_mean)/HOURLY_mean
Within Variability	SUM_Nvar	SUM_sd / SUM_Mean
	WINT_Nvar	WINT_sd / WINT_mean
	WD_Nvar	WD_sd/WD_mean
	WE_Nvar	WE_sd/WE_mean
	AM_Nvar	AM_sd/AM_mean
	PM_Nvar	PM_sd/PM_mean
	NI_Nvar	NI_sd / NI_Mean
Extremes	SUM_Nmax	(SUM_max - SUM_Mean) / SUM_Mean
	WINT_Nmax	(WINT_max - WINT_Mean) / WINT_Mean
	WD_Nmax	(WD_max - WD_Mean) / WD_Mean
	WE_Nmax	(WE_max - WE_Mean) / WE_Mean
	AM_Nmax	(AM_max - AM_Mean) / AM_Mean
	PM_Nmax	(PM_max - PM_Mean) / PM_Mean
	NI_Nmax	(NI_max - NI_Mean) / NI_Mean

The seasonal indicators are based on weekly averages rather than monthly averages for two reasons. Firstly, this provides a consistent link between the three different indicator types listed in Table 4.5. And secondly, weekly averages provide smoother patterns within seasons than do monthly averages.

Based on this methodology, twenty one indicators were calculated to highlight the most important aspects of variability in O_3 concentrations during different time periods. Prior to further analysis, correlations between the twenty one indicators were examined in order to detect any obvious redundancy in the variables. All variables showed moderate levels of correlation, and for some the Pearson correlation exceeded 0.80.

Table 4.5 Description of temporal indicators

	Seasonal	Weekday/weekend	Diurnal	Suggested metric	Application
Average	Normalised mean for: SUM WINT	Normalised mean for: WD WE	Normalised mean for: Am PM NI	Normalised mean: Meanp-Meany/Meany	Distinguishes between sites that have higher or lower average concentrations during the specified period compared to the average for the year.
Extreme	Normalised average daily max and min for: SUM WINT	Normalised average daily max and min for: WD WE	Normalised average daily max and min for: Am PM NI	Normalised average maxima: |Max|p-Meanp/Meanp Normalised average minima |Min|p-Meanp/Meanp	Distinguishes between sites that have higher or lower extreme concentrations during the specified period, relative to the average for the period (Note |….| depicts averages of the specified measures)
Within variability (i.e. within each period)	CV for: SUM WINT	CV for: WD WE	CV for: Am PM NI	Coefficient of variation: SDp/Meanp	Distinguishes between sites with more or less variable concentrations during the specified period, relative to the average for the period

Definition of time periods:
SUM: June to August WINT: December to February
WD: Monday to Friday WE: Saturday and Sunday
AM: from 06.00 to 12.00 PM: 13.00 to 19.00 NI: 22.00 to 05.00

4.2.2 Principal component analysis (PCA)

As noted previously, the correlations between some of the indicators was found to be high. Rather than taking an arbitrary decision to remove one of each pair of highly correlated indicators, which could lead to losing some information, further analysis is required to produce informative but uncorrelated components on the basis of the twenty one indicators.

There are two specific methods to achieve this task: PCA and factor analysis (Jolliffe, 2002). Both methods are used as a data reduction method, but the latter has the aim of revealing any latent variables within the observed or measured variables, whereas the former is used to derive a comparatively small number of variables which convey as much of the information in the observed or measured variables as possible (Leech et al., 2008). PCA is the preferred analysis here, as the purpose is to reduce the number of indicators, and to eliminate the correlations between variables used in subsequent analysis. Previous studies aiming to classify air pollution monitoring sites have followed the same scheme (i.e. using PCA, followed by cluster analysis). As noted previously, one such example is the study by McGregor (1996) which follow the approach of (Yarnal, 1992).

PCA is a multivariate technique designed to help understand the complex relationships among a set of variables, considered simultaneously, by using the correlations amongst them to create a set of components. PCA forms uncorrelated linear combinations of the observed variables. The first component explains the most variance, and successive components explain progressively smaller portions of the variance, and all are uncorrelated with each other. The method can be used when a correlation matrix is singular (Jolliffe, 2002), such as the case with the indicators derived in section 4.2.1.

The variables, in this case the indicators, first need to be examined to see if they conform to several statistical requirements in order to determine the appropriateness of PCA. The created indicators were therefore screened to test that the correlations between them justified the application of PCA. To be considered for PCA, all correlations should exceed 0.3; however, partial correlations (i.e. the correlations between indicators taking into account the effect of other indicators) should be small. This correlation assessment can be done using the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) tests. The KMO is the ratio of the squared correlation between indicators to the squared partial correlation between indicators (Field, 2009). The KMO statistic ranges between 1 and 0, where 1 indicates that correlations are compact and running the analysis will produce reliable

components. There are two values of KMO: one for multiple and the other for individual variables, both of which have to be greater than 0.5 (Field, 2009).

The next step was to determine how to select the appropriate number of principal components (PCs) to be extracted. Three criteria were chosen for this purpose:

1) Latent root criterion (eigenvalues): the eigenvalue is the variation explained by a component; any component with an eigenvalue greater than 1 is considered significant (i.e. has a significant amount of variation) (Kaiser, 1960).

2) Percentage of variance: based on achieving a particular cumulative percentage (e.g. established a priori as x percent) of the total variance extracted by consecutive components.

3) Scree plot: used to identify the optimal number of extracted components before the unique variance begins to dominate the common variance. A scree plot is a graph plotting each eigenvalue (Y axis) against the associated component (X axis) (Cattell, 1966). The point of inflexion of the curve is taken as the cut off point for the optimal number of components.

Once the PCs have been extracted, they need to be interpreted. Based on the un-rotated matrix, the first component depicts the best summary of the linear relationships between variables (indicators) and contains all variables with high loadings. The other components comprise the other variables, with lower loadings (Field, 2009). Interpretation of the components can be improved by obtaining the rotated matrix which eliminates the combined ambiguities in the un-rotated solutions. In other words, it reorganises the variance from early components to later components to achieve a more meaningful component model. Therefore, the rotation effectively maximizes the variable loading in a single component and minimizes the number of variables with high loadings in each factor.

Factor rotation means that the reference axes of the components are rotated about the original axis until another location has been reached. An orthogonal rotation is at 90 degrees, while an oblique rotation has no constraints and is more flexible, and allows components to be correlated. An orthogonal rotation is appropriate if the objective is to obtain a set of uncorrelated components needed for prediction techniques; this agrees with the aim, here, as these components will be used in a subsequent cluster analysis.

There are three approaches to rotation (varimax, quarimax and equamax) under orthogonal rotation, depending on the type of component desired. As noted, the objective here is to obtain a set of uncorrelated factors that are linear combinations of the initial variables, explaining most of

the variation in the data. Varimax is most appropriate in this situation because it maximizes the dispersion of loading within components, to produce more interoperable components compared to other approaches.

There is a widely used non-mathematical proposition for determining the significance of the variable loadings. As the component loading is the correlation between each variable and the factor, so the square of the loading is the variable's total variance accounted for by that component. By convention, loadings of 0.5 or more are considered significant (denoting 25% of the variance accounted for by the component); 0.4 is important; and 0.3 is not important since this means that the component explains only ca. 10% of the variable's total variation. Note that the square of the component loadings (correlation) reflects the total variation of the variables explained by the component; therefore, loadings should exceed 0.7 to account for 50% of the variance (Hair et al., 1998). Stevens (2002) says that only variables with a loading of more than 0.4 (denoting 16% of the variance accounted for by the component) should be considered appropriate for interpretation.

On examining the variables that load high on each component, a descriptive label or name is then assigned describing each respective component. The variables with highest loadings should have the most influence on the label of the component.

To be appropriate for PCA, both KMO measures for the sampling adequacy should be greater than 0.5. Here they were 0.9. This indicates that the twenty one indicators are appropriate to be used in PCA.

The scree plot (Figure 4.2) shows the change in the eigenvalue for each additional component. The point of the inflexion indicates that four components are appropriate, for at this level a stable plateau is reached.

Table 4.6 also shows that the fourth component is the last for which the eigenvalue exceeds 1. These four components combined represent 88.4% of the variance of the twenty one indicators. Furthermore the communalities in the un-rotated matrix are larger than 0.5, which indicates that a large amount of variance in all indicators has been extracted by the component solution.

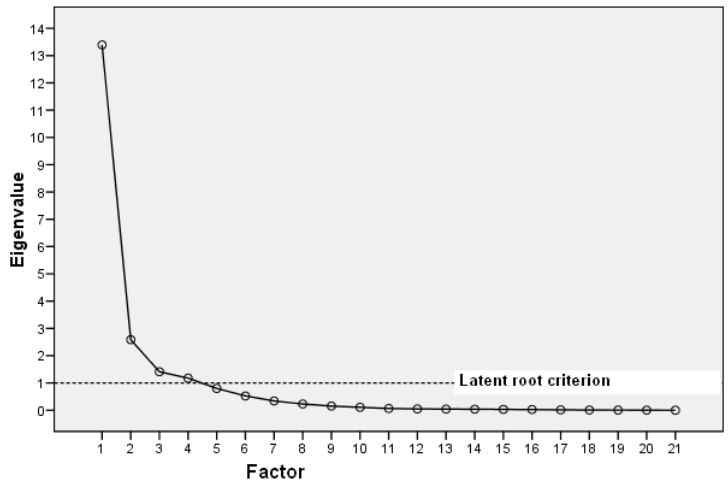


Figure 4.2 Scree Test for Component Analysis

Table 4.6 Results for the extraction of principal components showing the eigenvalues and the total variation explained by principal components

Extraction sum of squares loading				Rotation sum of square loading	
Components	Eigenvalue	% of variance	Cumulative % of Variance	% of variance	Cumulative % of Variance
1	13.39	63.76	63.76	25.59	25.59
2	2.58	12.32	76.08	24.11	49.70
3	1.41	6.71	82.80	23.21	72.92
4	1.17	5.59	88.40	15.48	88.40
5	0.80	3.81	92.21		
6	0.52	2.49	94.71		
7	0.33	1.61	96.32		
8	0.22	1.09	97.41		
9	0.15	0.72	98.14		
10	0.10	0.51	98.65		
11	0.06	0.30	98.96		
12	0.05	0.22	99.19		
13	0.04	0.19	99.38		
14	0.04	0.17	99.56		
15	0.03	0.15	99.71		
16	0.03	0.11	99.83		
17	0.01	0.06	99.89		
18	0.009	0.04	99.94		
19	0.006	0.03	99.97		
20	0.004	0.02	99.99		
21	0.000	0.0004	100		

The first component accounts for the largest amount of variance, with most indicators loading high on this component (63.8% compared to 12.3, 6.7, and 5.6). This suggests that Interpretation of the un-rotated components will not be easy. Therefore, a rotation of the matrix is needed to redistribute the variance between the components and make the interpretation more meaningful and simpler. After rotation, the first component accounts for only 25.6% of variance (compared to 24.1% 23.2% and 15.5% respectively in the other components).

In an effort to name each component, all components with high loadings from the rotated solution were used, bearing in mind that indicators with higher loadings influence the naming of the component to a greater extent (i.e. what the component actually represents). Table 4.7 shows the component solution derived from the analysis using the varimax rotation of the twenty one indicators of hourly O_3 concentrations. The cut-point for interpretation purposes was defined as all loadings above 0.4.

Table 4.7 VARIMAX-Rotated Component Analysis

Indicators	Principal Components			
	PC1	PC2	PC3	PC4
WD_Nmean	-0.86			
WE_Nmean	0.86			
PM_Nmax	0.83			
WE_Nmax	0.71			0.48
WD_Nmax	0.71			0.52
WINT_Nmean		-0.84	-0.40	
SUM_Nmean		0.84		
AM_Nmean		-0.75		
WE_Nvar		0.72		
WD_Nvar	0.40	0.68	0.41	0.43
AM_Nvar	0.52	0.61	0.54	
PM_Nvar	0.55	0.55	0.30	0.48
WINT_Nmax		0.54	0.53	
WINT_Nvar	0.44	0.50	0.45	0.43
NIGHT_Nmean			-0.97	
PM_Nmean		0.45	0.81	
NIGHT_Nmax	0.53		0.78	
NIGHT_Nvar	0.48	0.37	0.73	0.24
AM_Nmax	0.50	0.52	0.61	
SUM_Nmax				0.91
SUM_Nvar				0.89

Component 1 (PC1) seems to comprise monitoring sites with a contrast in average concentration between weekend and weekday. Also it includes sites with higher maximum concentrations in the weekday/weekend and afternoons. This contrast is likely to be produced from differences in emission intensity from human activities during the two periods of time. Variations in traffic density, especially, are implied, for this is the major source of NO (O_3 scavenger). PC1 is thus termed as a weekday/weekend component.

Component 2 (PC2) tends to represent sites having a contrast in average concentration between summer and winter. These sites also have high variations in concentration during the week and during the day. Interpretation of this component needs to be done with care, as the summer/ winter contrast is likely to reflect broad geographical influences of climate, which in turn are likely to relate to latitude and degree of continentality (coastal versus inland sites). PC2 is thus termed as a summer/winter contrast component.

Component 3 (PC3) captures sites having a contrast in the average concentration between night-time and afternoon. This is likely to be related mainly to the natural pattern of photochemical activity between day and night. It will thus help to distinguish areas in which there is a strong afternoon peak and low concentrations at night (e.g. due to the absence of any inter-regional transport of O_3). This PC is termed the afternoon/night contrast component.

Finally, component 4 (PC4) seems to include sites with high maximum concentrations and variation during the summer. This could highlight the effect of seasonal photochemical reactions. For example, sites located in warm, inland areas and far from any source of scavenging of O_3 will have very high O_3 concentration, especially during warm, sunny periods of the year. This PC4 is termed the high summer variation component.

As the names given to the PCs implies, each one highlights the importance of a different environmental factor on O_3 formation and destruction. The four most meaningful components will be used to classify the 1211 O_3 monitoring sites over Western Europe using hierarchal cluster analysis, as described in the following section.

4.2.3 Hierarchal cluster analysis (HCA)

The four PCs (based on the twenty one indicators) defined in the previous analysis were used in cluster analysis to classify sites into a number of types based on their temporal O_3 signature. There

are two main approaches: HCA and K-means clustering. The latter approach is mainly appropriate where the initial cluster centres (cluster seeds) need to, and can, be defined in advance. The process works iteratively, starting with the first seed. All cases within a pre-specified threshold distance of the seed are assigned to the associated cluster. Another cluster seed is then assigned and cases assigned in the same way. These steps continue until all cases are assigned to the predefined number of clusters. A major problem with K-means cluster analysis, however, is how to select the initial cluster seeds. The results of the analysis also depend on the order of the cases in the data set. Furthermore, the optimum number of clusters is often not known in advance. On the other hand, HCA can be run when the number of clusters is undefined and when the analysis is largely free of K-mean constraints (Willett, 1988). It thus provides a more powerful grouping approach and yields better quality clusters (Flemming et.al., 2005; McGregore, 1996; McGregor and Bamzelis,1995). Therefore, HCA was used in this analysis.

HCA is a multivariate technique used to group cases based on their shared characteristics, with the aim of maximizing variation between the resulting groups and minimising the variation between cases within each group. HCA searches for an underlying structure of cases (sites in this context) through an interactive process, using either agglomerative methods (the most common approach) or divisive methods to assign cases to a cluster, case by case, until all cases have been processed (Steinbach et al., 2003).

Agglomeration starts with all sites in separate clusters, then, based on the selected similarity measures, these are progressively joined into fewer, larger clusters, until all sites end in one inclusive cluster. Various measures may be used to assign cluster membership, including the correlation matrix between variables and distance measures. Correlation measures are rarely used because clustering depends on the magnitude of cases, not just the pattern of association between them. A distance measure is a measure of dissimilarity across the cases, and is converted to similarity by an inverse relationship. Euclidean distance, for example, is often used. This is the distance between two points represented by the length of the hypotenuse of a right angle triangle.

Various methods for clustering can be used, each based on different ways of defining between and within group variation. Ward's method is one of the most widely used and, perhaps, most robust methods of clustering (Henne et al., 2010, Flemming et al., 2005). This uses the sum of the squared distance between two clusters summed over all variables. The centroid method uses the distance between the centroid of two clusters. Both Ward's and the centroid methods use the squared

Euclidean or simple Euclidean distance, which is not affected by outliers. For this reason, Ward's method was selected in this analysis.

The agglomeration chart produced during HCA depicts the changes in the coefficient at each stage of the clustering process. Small changes in the coefficient indicate that relatively homogenous clusters are being merged, whereas large changes occur when two distinctly different clusters are merged. This can be used to inform the decision about the number of extracted clusters.

As already noted, the data used in HCA consisted of the 1211 O_3 monitoring sites (as the cases to be classified) and the 4 PCs produced from PCA, representing hourly O_3 patterns. The squared Euclidean distance was chosen as the similarity measure, using Ward's method.

Using the four components derived from the PCA in the cluster analysis produced the agglomeration schedule and coefficients depicted in Figure 4.3. From this scree plot, it appears that there are two possible cut points, where the curve flattens: at thirteen and seven clusters. Analysis of variance (ANOVA) was run for both sets of clusters to explore the extent to which the PCs explained the variation between the different site types. This showed that 67% of the variation between the groups was explained by the four PCs at the thirteen cluster step, and only 50% at the seven cluster step. Therefore, the cut point at thirteen clusters was selected, giving thirteen different site types for further analysis.

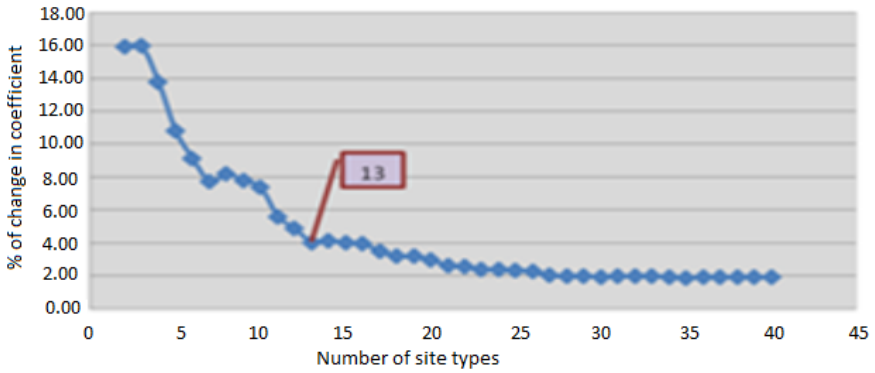


Figure 4.3 Scree plot for the HCA

Each group was next explored to see how the means of the PC scores vary within different groups. Results are shown in Table 4.8.

Table 4.8 PC scores for the 13 site types

Site type	Weekday/weekend contrast	Summer/winter contrast	Afternoon/night contrast	High Summer variation
1	-1.1	-0.3	-0.9	1.0
2	-0.4	0	-0.8	-0.8
3	-0.5	-0.7	0.8	-0.1
4	0.1	0.1	0	0.2
5	-0.1	-0.2	0.1	0.2
6	0.2	-0.3	2.3	-0.6
7	2.2	0.1	-1.3	-0.5
8	1.1	-0.8	0.1	0
9	0.6	0	0.1	1.6
10	-0.4	0.9	1.3	0.5
11	0.6	0.6	0.3	-1.7
12	-1.1	-1.1	-1.1	-0.5
13	-0.8	3.0	-0.2	0

The following points may be adduced:

- The weekday/weekend contrast distinguishes between site types 7 and 8 (with highest positive scores) and site types 1 and 12 (with highest negative scores).
- The summer/winter contrast distinguishes site types 13 (with very high positive score = 3) and other groups with high positive scores (site types 10 and 11) from those with high negative scores (site types 12, 8 and 3).
- The afternoon/night contrast distinguishes between site types 6 and 10 (with high positive scores) and site types 1, 7 and 12 (with high negative scores).
- The high summer variation component distinguishes between site types 1, and 9 (with high positive scores) and site types 11 and 2 (with high negative scores).

To understand whether the site types tend to associate with specific countries or are distributed in relation to environmental conditions, the different site types were mapped, as shown in Figures 4.4 to 4.6.

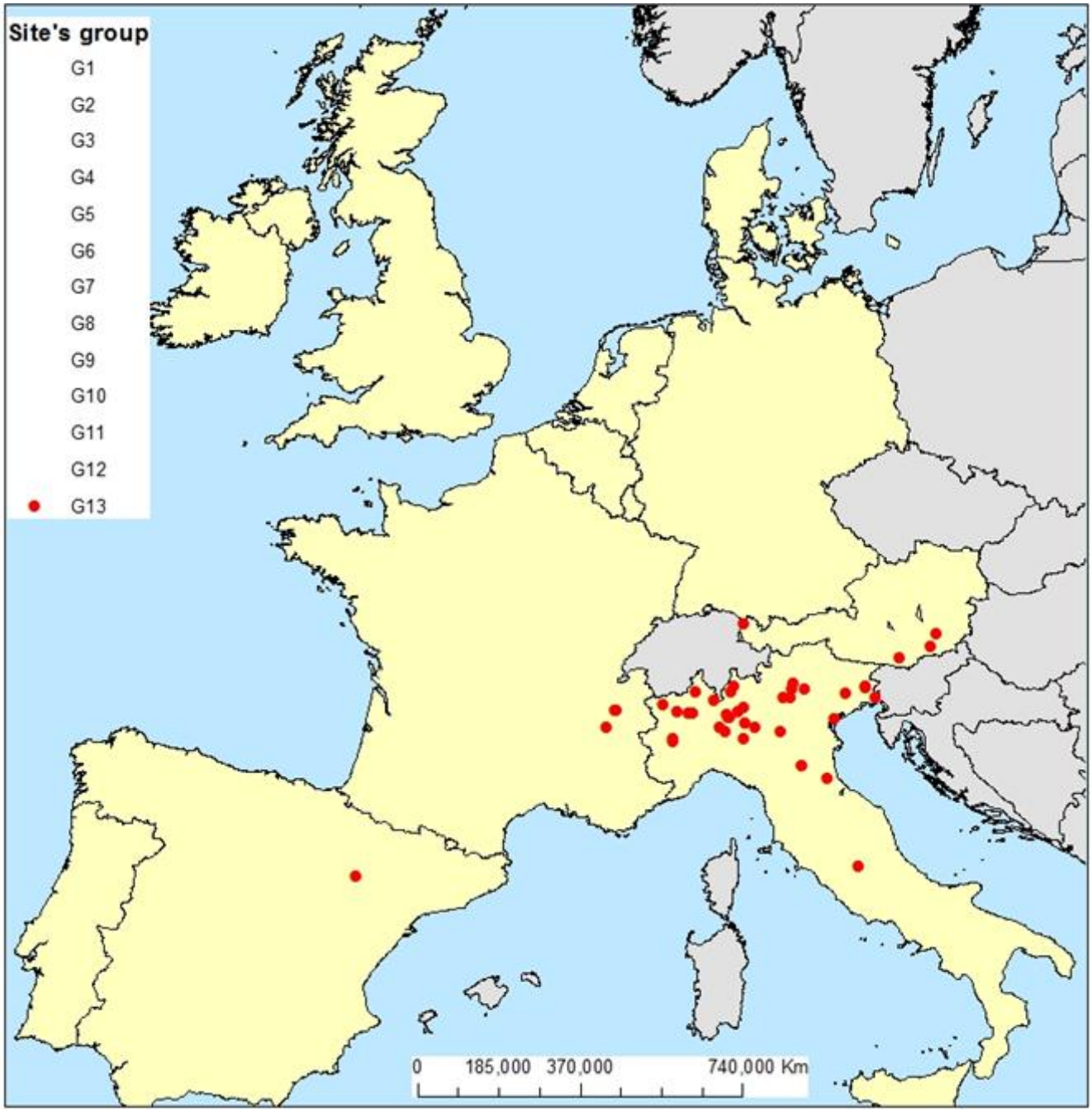


Figure 4.4 The distribution of site type G13

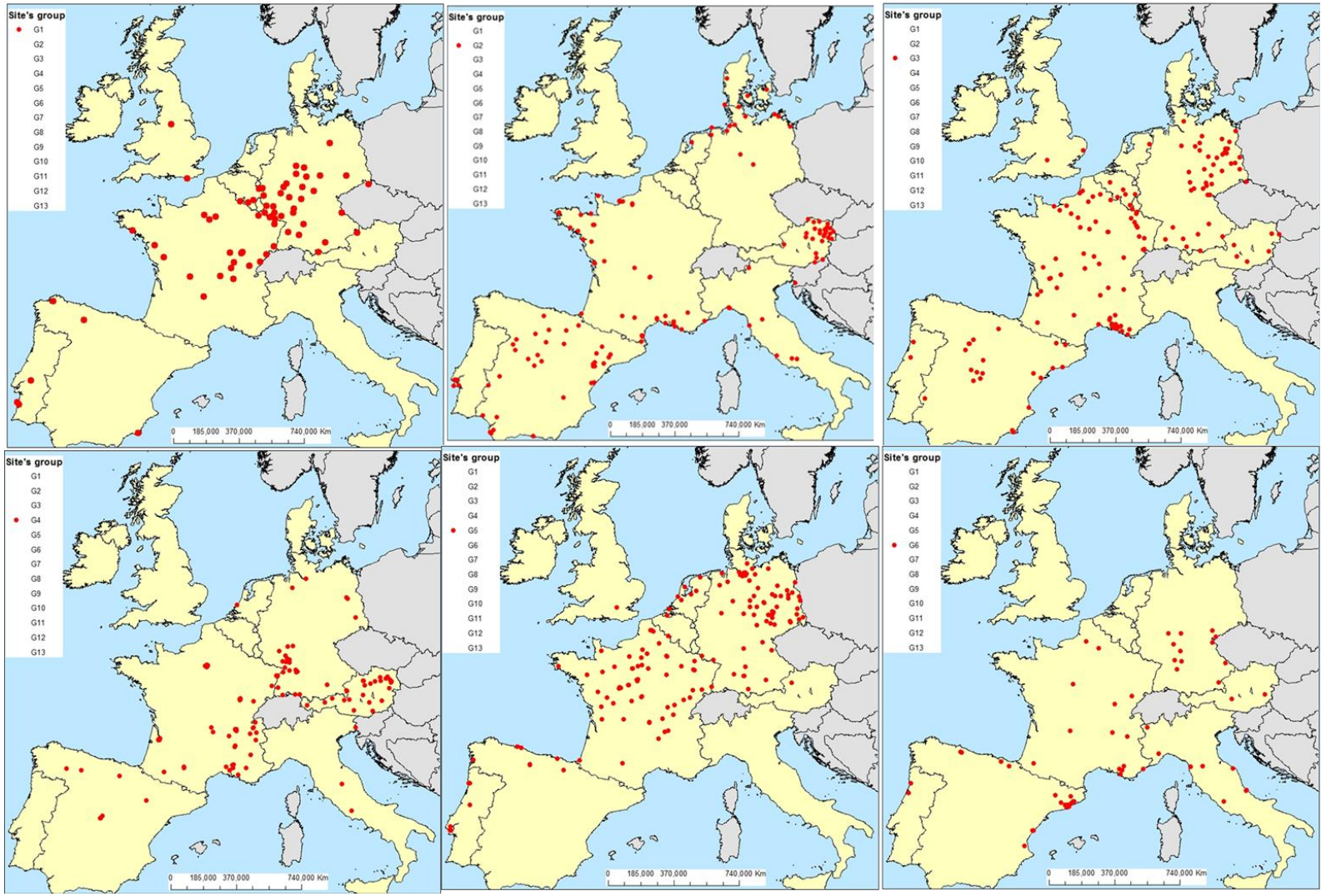


Figure 4.5 The distribution of site types G1 to G6

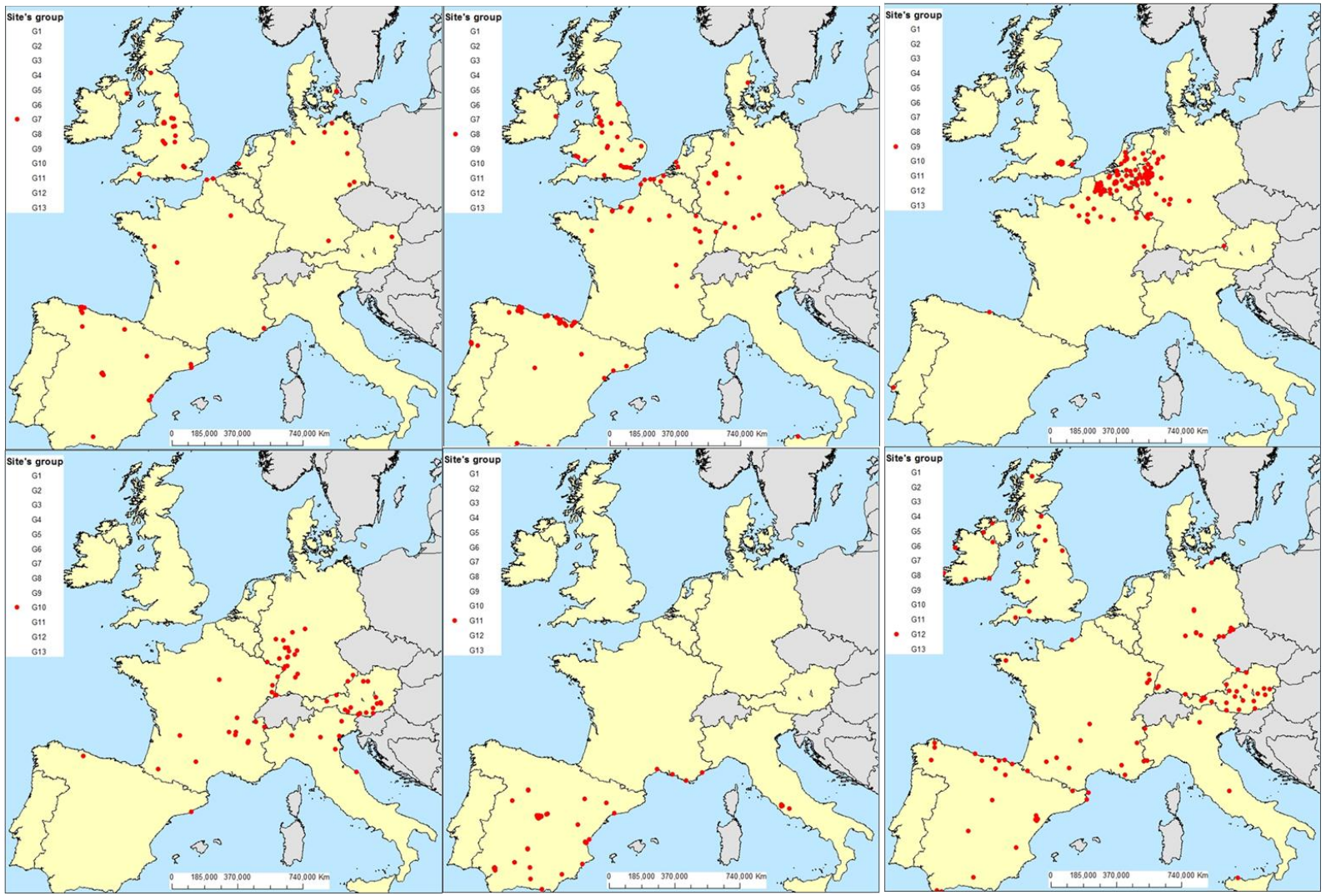


Figure 4.6 The distribution of site types G7 to G12

All site types are seen to be widely distributed, with the exception of site types 11 and 13 which are located mostly in southern Europe (the former mainly in Spain and the latter in Italy), and site type 9 which is concentrated in northern Europe.

Most countries include most of the site types: Spain and France, for example, have all site types while German contains all except site types 11 and 13. A few, however, show a more restricted range. The Netherlands has 16 sites of type 9, but none of site type 1, or 10 to 13 as illustrated in Table 4.9. Belgium, Denmark, and Ireland have only 4, 3, and 2 site types, respectively.

Table 4.9 Number of sites by site type in each country

		Site type													
EU country	**Location in study area**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Total**
AT	**Central**	2	27	6	25	1	2	1			17		22	4	107
BE	**Central**	3		6					3	22					34
DK	**Northern**		3					2	1						6
ES	**Southern**	3	36	22	7	7	20	29	34	1	2	46	24	1	232
FR	**central and southern**	24	40	61	53	59	14	6	18	37	19	4	16	3	354
GB	**Northern**	2		2		1		18	22	8			10		63
GE	**Central northern and**	30	13	43	27	67	12	8	15	34	19		14		282
IE	**Northern**								1				5		6
IT	**Southern**		10		3		8		1		6	5	4	35	72
NL	**Northern**		1	1	1	8		1	2	16					30
PT	**Southern**	3	7	2		6	2		4	1					25
Total No.		67	137	143	116	149	58	65	101	119	63	55	95	43	1211

The box plot in Figure 4.7 shows the long term average O_3 concentrations for sites in each site type. It is clear that site type 12 has the highest O_3 concentration, and also shows the highest variation. Site type 7 has the lowest mean concentration. The box plot also shows the presence of a large range between maximum and minimum concentrations, and a number of outliers in each site type. The interquartile range is specially high at site types 11 and 12. Overall, the box plot demonstrates that sites do not fall into distinct types in terms of their long-term mean concentration.

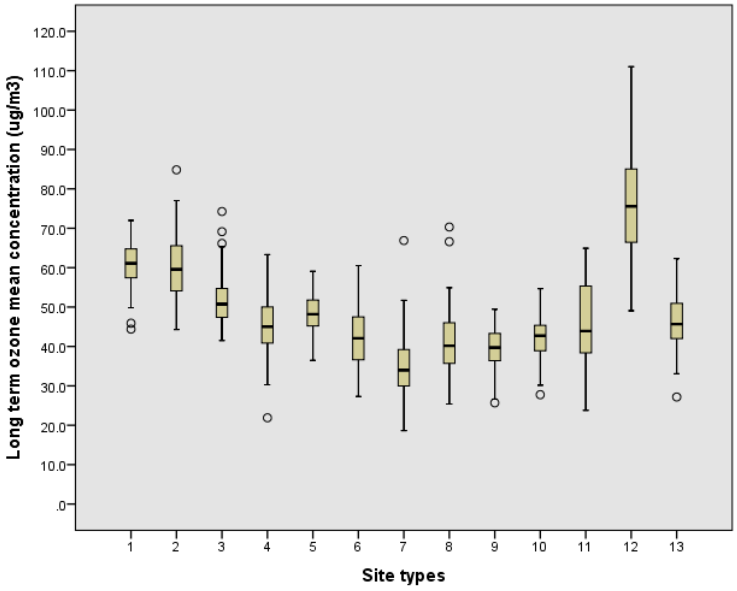


Figure 4.7 Box plot of the long term mean O_3 concentrations for sites in each site type

As noted, the site type classification needs to be assigned to unmonitored locations, as a basis for modelling the temporal patterns of O_3. For this to be done, an association needs to be found between the site type classification and a set of environmental variables, that can be measured at unmonitored locations. It is also helpful to name the site types, according to their environmental characteristics. Table 4.10 therefore shows the mean values, and 25th and 75th percentiles, by site type for a number of environmental variables across all 1211 sites. Group means were compared against the overall mean and percentiles to determine which environmental variables were influential, and from this an indicative name was suggested. This is reflected in the red/green colours in Table 4.10: red font highlights group means higher than the overall mean, while the red circle highlights variables and site types for which the mean exceeds the 75th percentile for the data as a whole. Green font shows cases where the site type mean is less than the 25th percentile of the data as a whole.

The variables used in this comparison relate to:

- The long term solar radiation, which reflects the intensity of photochemical processes; solar radiation increases from north to south so this is a good indicator of broad site location;
- Topography – represented in terms of altitude (m), topex (m), and distance to sea (km), all of which have impacts on O_3 concentration through their influence on temperatures, solar radiation, mixing and turbulence, and dispersion. Indirect associations also tend to exist

with the intensity of local emissions of precursor pollutants, since areas of high altitude and high relative relief (positive topex values) tend to be less suitable for urbanisation, and thus have low road densities and few industrial sources.

- Extent of urban land cover, specifically in terms of high density residential, low density residential and industrial/commercial land within a 1Km window size. Together, these indicate the intensity of urban activities and the density of population in the surrounding area.
- Rural land cover, represented in terms of the area of forest and agriculture within 1Km window radius size.
- Road density, expressed in terms of the total road length in metres within a 300m window radius size, which indicates intensity of pollution from road traffic emissions, but also provides an indirect indication of urbanisation.

Based on these variables and consideration of their geographic distribution (Figures 4.4 and 4.6), the following indicative names are assigned to the site types:

- Site type 1 has a high mean area of forested land, is relatively far from sea, has high altitude and topex values, and only small amounts of built-up land. Together these indicate sites comprising inland hills with little infrastructure. Taking the most important characteristics, the suggested name is 'Forested hill-lands'.
- Site type 2 comprises mixed land use areas, with about 47% of the land area under urban uses and 34% rural; and with high long levels of term solar radiation. The suggested name is 'Sunny mixed-use'.
- Site type 3 likewise has mixed land use with about 46% of forest and agriculture, and 39% urban area, but topex is low(mean=-12m). This is thus named 'Mixed use moderately sheltered'.
- Site type 4 has a high road length (mean ca. 1410 metres), with 75% of the area classified as urban, and with low topex values (mean = -11m) indicating that the sites are in sheltered areas. The distance to sea implies inland areas. The suggested name is 'Urban inland moderately sheltered'.
- Site type 5 comprises 62% urban area areas, and the distance to sea implies a somewhat inland distribution (though as the map in Figure 4.4 shows, this is not marked). Solar radiation is rather low (the second lowest of all site types). The indicative name given is 'Urban inlands'.

- Site type 6 is characterised by strongly sheltered locations, as shown by the strongly negative topex value (-27m, which >-20m). Land use is mixed with 43% urban and 44% rural areas; sites are centred in southern and central Europe. The suggested name is 'Sunny mixed use strongly sheltered'.
- Site type 7 comprises flat areas with high densities of urban land (80%) and heavy traffic loads (as indicated by the high value for roads within 300m). Thus the indicative name given is 'Heavily trafficked urban'.
- Site type 8 is broadly similar to site type 7 but the extent of urban area is somewhat lower (60%), as are traffic densities. Solar radiation is high, and both altitude and topex values relatively low. It also has the lowest value for distance from the sea, implying a maritime climate, as is also shown by Figure 4.4. The suggested name is 'Maritime urban moderately sheltered'.
- Site type 9, like site type 8, has moderately high urban extent (63%) and low altitudes (the lowest of any site type). Sunshine levels are also low, implying that it is located in northern Europe. The indicative name is 'Northern urban'.
- Site type 10 comprises inland strongly sheltered locations, as shown by the very low topex value of -28 metres and the relatively high distance to the open sea (261 km); moderate urban densities are indicated by the land cover data, with a substantial area of rural land cover. Therefore, the suggested name is 'Inland populated strongly sheltered'.
- Site type 11 is located mainly in southern Europe and has a climate characterised by high levels of solar radiation (mean = 168 w/s). High density urban areas make up 67% of the land area, and road density is relatively high, as is the area of industrial/commercial land; agriculture makes up only land 15%. Altitude is moderately high (377 metres). The suggested name is 'Southern urban uplands'.
- Site type 12 comprises high altitude areas (mean 789 metres), with high relative relief (topex = 39m) occupied by forest areas. Thus the suggested name is 'Forested mountains'
- Site type 13 is interpreted as low-relief lowlands (sheltered) located in the south of Europe, with moderate urban densities 63%; therefore the suggested name is 'Southern populated strongly sheltered'.

Table 4.10 Environmental descriptive classification for site types with suggested name

Site type	Sun radiation	Altitude	Topex	Dis2sea	Urban 1000	TR 300	FOREST 1000	AGR 1000	**Rural 1000**	Suggested name
1	122	424	21	287533	18	484	42	22	**64**	**Forested hill-lands**
2	144	253	3	129219	47	874	9	25	34	**sunny mixed use**
3	129	244	-12	216391	39	730	16	30	46	**Mixed use moderately sheltered**
4	127	259	-11	262362	75	1410	5	10	15	**Urban inland moderately sheltered**
5	121	148	-3	218069	62	977	4	21	25	**Urban inland**
6	145	210	-27	159656	43	866	15	29	44	**Sunny mixed use strongly sheltered**
7	135	235	0	143488	80	2344	1	4	5	**Heavily trafficked urban**
8	133	113	-10	124193[A]	60	1366	3	15	18	**Maritime urban moderately sheltered**
9	114	70	-2	164572	63	1089	4	20	24	**Northern urban**
10	124	275	-28	261120	58	1093	9	22	31	**Inland populated strongly sheltered**
11	168	377	-2	172935	67	1909	1	15	16	**Southern uplands**
12	133	789	39	168817	5	282	41	15	56	**Forested mountains**
13	134	228	-20	145449	63	1002	9	20	29	**Southern populated strongly sheltered**
Mean	131	264	-3	189766	42	1057	12	20	32	
25%	114	44	-11	68057	0	255	0	0	0	
75%	154	383	6	303725	84	1595	13	31	44	

[A] an indication of maritime location as it is the lowest value across the thirteen site types

Red font= value >mean circle ≥ 75% Green font= value ≤ 25%

Sheltered if topex has a negative sign moderately sheltered when topex ≥ -10m and <-20m strongly sheltered when topex > -20m

Exposed if topex has a positive sign moderately exposed when topex ≥ 20m strongly exposed when topex > 30m

Environmental factor: Sun radiation (long term solar radiation), Des2sea (Distance to sea), Urban_1000= High density residential low density residential+ and industrial/commercial lands, FOREST_1000 (forest), AGR_1000 (agriculture), RURAL_1000= forest+ agriculture, and TR_300 (total road length).

4.2.4 Defining site types in term of Environmental factors

In the previous section it was shown that the 1211 O_3 monitoring sites could be classified into 13 site types using temporal indicators (i.e. the four PCs) in a HCA. The purpose of this section is to establish a relationship between these predefined site-types and a set of environmental variables for the monitored locations by deriving discriminant functions. These functions will later be used to predict the probabilities of site-type membership for unmonitored locations.

4.2.4.1 Selection of environmental predictors

Classifying unmonitored locations is vital, since the lack of O_3 data at these locations means that they cannot be assigned to site types using indicators as before. Instead, site type membership has to be imputed using exogenous variables, for which known associations can be deduced at sampled sites (e.g. relating to emission sources/intensity, dispersion characteristics, atmospheric chemistry etc).

As already stated, site type was originally classified on the basis of a series of indicators of the temporal variability of O_3 concentrations. The first requirement, therefore, is to find environmental predictors that can be shown to correlate with the site types, and thus provide a means of predicting membership.

Predictor variables were selected as proxies for the most important determinants of O_3 concentrations: emissions to the atmosphere (roads type, land cover data), the physical impact of landscape (topographic features), and meteorological factors. Relationships between these variables and the site types were analysed using MLOR. Table 4.11 lists the rationale of the variables included.

Multicolinearity has to be considered in MLOR (Field, 2009, Kempen et al., 2009). A correlation matrix was therefore constructed to identify very highly correlated variables ($R>0.8$). In these cases, one variable of each highly correlated pair was omitted. Choice between the correlated variables took account of the radius of the window on which they were based, in order to ensure that the variables retained reflected both the local effects (100m to 1000m) and regional effects (5Km and 10Km). The retained variables are shown in Table 4.12.

Table 4.11 The rationale for variables include in MLOR

Environmental variable	Formation and dispersion
Road types:	Proxy of traffic impact
Major	Source of O_3 precursors
Secondary	Differentiates between traffic sites and other types
Local	
Residential area lands:	Proxy of human activities
High	Source of O_3 precursors
Low	Differentiate between urban and suburban sites and other types
Industrial/commercial lands	Proxy of human activities
	Source of O_3 precursors
	Differentiates between industrial and other types
Forest land	Source of O_3 precursors
	Differentiates between remote area and other types
Green lands:	
Agriculture	Source of O_3 precursors and dispersion surface
Herbaceous	Differentiates between rural and other types
Topography:	Proxy of physical impact of topography in O_3 formation and transportation.
Distance to sea	Proxy of maritime in O_3 by distance to sea.
Altitude	Differentiate between places in terms of their topography (absolute and relative altitude)
Topex	
Meteorological factors:	Reflect the role of metrology in temporal variation of O_3 and on all above variables.
Temperature	Proxy of the active photochemical process of O_3 by Temperature.
Wind speed	Proxy of the transportation of O_3 by wind speed.
Total precipitation	Proxy for wet deposition of O_3 by total precipitation.

Table 4.12 MLOR predictor variables

Variables	Abbreviation	Predictor variable	Window's radius size (m)	unites
Land cover	Highdr	High density residential land	**500,1000, 10000**	Percentage
	Lowdr	Low density residential land	**1000, 10000**	Percentage
	Ind/Com	Industrial/commercial	**1000, 10000**	Percentage
	Herb	Herbaceous land	**1000, 5000**	Percentage
	Agri	Agriculture land	**1000, 5000**	Percentage
	Forest	Forest land	**1000, 10000**	Percentage
Topographical	Dis2Sea	Distance to sea		Kilometre
	Altitude	Altitude (height above sea level)		Metre
	Topex	Topex		Metre
Road length	MR	Motorways	**100, 500,10000**	Kilometres
	SR	secondary Roads		Kilometres
	LR	Local Roads		Kilometres
Meteorological	TP	Winter and summer average total precipitation		mm
	TEMP	Winter and Summer average temperature		C^{o}
	WS	Winter and summer average Wind speed		m\s

4.2.4.2 Regression analysis

As mentioned in section 4.1, two main methods can be used for the purpose of classification: MLOR and DA. They are used to predict a categorical variable (e.g. class membership) from a set of continuous values or/and categorical values (predictor variables). Both methods define the relationship between multiple independent variables and the categorical dependent variable by forming a combination of the independent variables (Leech et al., 2008). However, differences exist in the underlying assumptions. The statistical assumptions required for DA are as follow:

- The independent variables have a multivariate normal distribution.

- Homogeneity of variance and covariance for all independent variables across all groups.

Common features of the data used for studies such as this are the existence of strong correlations among the independent variables and outliers in the independent variables, which can alter the results, and mean that these assumptions are not met. MLOR, unlike DA, has no such assumptions (Menard, 2002, Homer & Lemeshow, 1989 cited in Field 2009), and it performs well compared to DA (Morgan et al., 2003); for this reason it is often preferred. Both methods can also provide estimates not only of the most likely class of each unmonitored location, but also the probability of membership of the different classes. As has been clear from the descriptions of the different site types, there were no discrete cut points between the different classes, and in reality they overlap each other to a considerable degree. For this reason, it was considered more appropriate to take account of these probabilities rather than 'forcing' every unmonitored site into a single site type. This was done by using the probabilities as weights to combine time functions for each site type into a 'best-estimate' of the time functions at each site.

MLOR predicts a categorical dependent variable Y using a combination of X predictors (categorical and/or continuous) each multiplied by its respective regression coefficient. MLOR is thus similar to linear regression, except that in multiple regressions dependent and independent variables have to be continuous. The MLOR model can thus be represented as in Equation 4-1:

$$G_i = \frac{Pi}{1-Pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + \varepsilon \qquad \textbf{Equation 4-1}$$

where G_i is a logarithmic function of the probability of being a member of site type (P_i) divided by the probability that it is not ($1-P_i$) relative to the reference site type; β_0 is the G intercept, and β_1 to β_k are the coefficients of the covariates (predictors), measuring the contribution of X_1 to X_k, respectively. K is the number of covariates and ε is the difference between the observed and predicted value, i.e. error.

In this study there are twelve equations, one for each of the site types defined by the predictor variables, relative to a reference site type. The twelve equations can be used to calculate the probability (*G*) that a site is a member of each of the twelve groups. The value of *G* for the reference site type is equal to zero.

MLOR predicts the probability of the G_i *(*categorical) value occurring, on the scale between 0 and 1 (Field, 2009), so:

$$P(Gi) = \frac{EXP(Gi)}{EXP(G1)+EXP(G2)+\cdots EXP(Gn)}+\varepsilon$$ **Equation 4-2**

where *P(Gi)* is the probability of site type *i*, *n* is number of site types predicted by dividing the *G* of site type *i* to the sum of all other site types *G*n.

Therefore, the exponent EXP (β), known as the odds ratio (OR), is the crucial value in interpretation of MLOR. It indicates the proportion by which the odds of the *Gi* changes when the predictor variable is changed by one unit with respect to the reference category (Menard, 2002; Hosmer & Lemeshow, 2002, cited in Field 2009). A value of EXP (β) greater than 1 indicates that the probability of occurrence increases due to an increase in the values of the predictor variable. A value of EXP (β) less than 1 indicates that the probability of occurrence decreases due to an increase in the value of the predictor variable.

As Field (2009) suggests, model evaluation can be done as follows:

- Observed and predicted values have to be compared, using the log-likelihood, which is a sum of probabilities associated with the predicted and observed outcomes. Therefore, comparing log-likelihoods involves comparing the intercept (as baseline) with the log-likelihood using the combination of predictors. If the log-likelihood decreases, it can be assumed that the model is improved by adding more predictors.
- Cox and Snell (R^2_{cs}) and Nagelkerke's (R^2_N), coefficients are used to measure the significance of the final model on a scale between 0 and 1, equivalent to R^2 in multiple linear regressions. A value close to 1 indicates that the model predicts the outcome more-or-less perfectly.

Two methods for including the predictor variables can be used in MLOR (Field, 2009): enter (i.e. the selected predictors are entered individually or as a group, based on the decision of the user) or stepwise (i.e. predictors are entered on the basis of specific statistical criteria). The former approach is more appropriate for theory testing; the latter is appropriate for exploratory work aiming to find a model to fit the available data.

The stepwise technique can be applied in either a forward or backward mode, as demonstrated by Field (2009). In the former method, the most significant (typically $P \leq 0.05$) predictors are sequentially added to the model until none of the remaining predictors is significant; each predictor is also examined at each step to see whether it should be removed, using the likelihood ratio statistic. The backward method begins with all the predictors in the model, and then predictors are removed at each step, if they do not have a significant effect on the model fit to the data. Using the backward

method can produce complex models, while the forward method might miss important and potentially significant variables. For this analysis a forward stepwise approach was preferred, since this was likely to generate more conservative, and simpler, models.

As a forward stepwise approach was selected to run the MLOR analysis, a supervised method was used, in which the environmental predictors in Table 4.12 were entered in a predefined sequence reflecting their order of importance, as shown in Table 4.11. In each case all the selected window radius sizes for that variable were entered at each step. In the first step, major roads were entered and a major roads model was built using all buffer sizes that gave significant contributions. All the major roads windows that were not included were then dropped from the analysis. Next, secondary roads were entered along with the predictors from the previous step (i.e. the major roads model). A major + secondary roads model was thus built. All the roads variables that were not included in the major + secondary roads model were then dropped from the analysis. The same procedure was repeated for the remaining land cover predictors, in their order of importance: i.e. local roads, high density residential, low density residential, non-residential, forest, agriculture and herbaceous land covers. Then the topographic variables were entered in the same way, and finally the meteorological factors, in sequence.

Assessment of multicolinearity was done after each stage using the variance inflation factor (VIF). This was performed by running a separate multiple regressions using site type as the dependent variable and the predictors as independent variables, since MLOR cannot provide a VIF.

MLOR was carried out in SPSS v15, using the thirteen site types produced from HCA in Phase1 as the dependent variable, and the environmental variables as independent variables. The logistic regression (NOMREG) module of the SPSS was chosen since the dependent variable had 13 categories (site types). Site type 5 was defined as the reference category as it has the greatest number of sites.

In this analysis, the probability of the model chi-square being significant was 0.0000 ($P \leq 0.05$). The null hypothesis that there was no difference between the model without independent variables and the model with independent variables was therefore rejected. Table 4.13 shows that inclusion of the significant predictor variables decreases the log-likelihood from 6034 to 3436. The existence of a relationship between the independent variables and the dependent variable was therefore supported. R^2_{cs} and R^2_N were equal to 0.9, which indicates that the overall accuracy of the model is good.

The relative significance of the variables in predicting the site type membership broadly reflects their importance in determining O_3 concentrations. For example, considering the first seven predictors, as ordered in Table 4.13, it can be seen that:

- Local roads (i.e. road density within 10Km window size) are important predictors, and indicate the importance of these as sources of O_3 precursors.
- Wind speed, which is seen to be important in discriminating between site types, likewise plays an important role in transport and mixing of O_3.
- Altitude reflects how distributions of O_3 concentration vary in the vertical dimension.
- Summer temperature is important because it influences photochemical reaction, and is also very highly correlated with solar radiation (R>9.0). It thus acts to distinguish between sites in different regions (northern vs. southern or coastal vs. inland).
- Distance to sea represents the maritime effect (i.e. transported O_3 from sea).
- Residential land follows in the order of importance, and provides indicators of emissions of precursors from human activities.

Table 4.13 Summary of the MLOR statistics for final model

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	6034.472			
LR_10000	5555.223	479.248	12	0.00
WS_win	5089.377	465.846	12	0.00
Altitude	4753.574	335.803	12	0.00
TEMP_sum	4439.226	314.349	12	0.00
Dis2sea	4236.574	202.652	12	0.00
Lowdr_1000	4101.383	135.191	12	0.00
Highdr_1000	3991.602	109.781	12	0.00
TP_win	3890.966	100.636	12	0.00
Topex	3816.509	74.457	12	0.00
SR_10000	3750.088	66.421	12	0.00
Herb_5000	3686.381	63.707	12	0.00
Forest_10000	3642.255	44.125	12	0.00
Ind/Com_1000	3607.325	34.93	12	0.00
Agri_5000	3572.916	34.409	12	0.00
Highdr_10000	3540.698	32.218	12	0.00
SR_100	3512.403	28.294	12	0.01
MR_10000	3486.061	26.343	12	0.01
Forest_1000	3460.115	25.946	12	0.01
Lowdr_10000	3436.934	23.181	12	0.03

The MLOR model presented in Table 4.14 provides the odds ratio of probabilies of membership of each site type, as well as the significance of each predictor to the overall model. This enables the effects of a change in the predictor variables to be estimated. For example, an increase of 1 km in local road length within a 10Km window radius size would increase the probability of occurrence of site type 4 by 14% (OR=1.14), site type 6 by 21%, site type 7 by 44%, site type 8 by 25%, and site type 9 by 26%, compared to the reference site type 5.

Table 4.15 presents a confusion matrix, showing the probability that sites will be placed in the wrong site type by the MLOR (relative to their initial classification by the HCA). About 15% of sites in site type 1, for example, will be incorrectly assigned to site type 12 (and thus be allocated a temporal model for that site type), while 13% will be assigned to site type 3. Likewise, 26% of type 3 sites will be given a site type 5 model, while 15% will be modelled as site type 2. Examination of the matrix suggests that, in particular, confusion occurs between site types 1, 3, and 12, between site types 7, 8, and 9, and between site types 10 and 4. The overall percentage of sites given the same classification as in HCA was 52%.

These errors highlight the fact that the site types overlap and are not distinct, and emphasise the importance when modelling of using the probabilities of site type membership to weight the models at any specific location.

4.3 Summary

This chapter has shown that a single space-time model of O_3 concentrations will not be effective across Western Europe unless the temporal variations in concentrations are properly characterised. Although it is often assumed that pollutant concentrations at different sites vary in harmony (at least within a specific geographic area), the results obtained in this study show that sites within Western Europe, categorised into thirteen site types, behave somewhat differently, depending on their context.

The thirteen site types are not discrete and site types defined both from HCA, using the temporal component indicators, and from MLOR, which aimed to find the relationship between the site type and the exogenous environmental variables, have uncertainty. The approach taken in subsequent analyses will thus be to apply the site type classification to sites probabilistically.

Temporal models will be created for each site type, using data from the monitored sites. The probabilities of site type membership will then be estimated at each location, and these used to weight and combine the different models to predict temporal variations in O_3 concentrations. This will be described and discussed in Chapter 6. The next chapter explains the spatial modelling that is done to estimate the underlying geographic variations in long term average concentrations.

Table 4.14 The estimated odds ratios EXP(β) of the final MLOR for O_3 site types in Western Europe

Predictors	Site type1	Site type2	Site type3	Site type4	Site type6	Site type7	Site type8	Site type9	Site type10	Site type11	Site type12	Site type13
Dis2sea	1.001	0.996*	1.001	1.004*	1	0.994*	0.996*	1.001	1.005*	0.993	0.993*	1.017*
Highdr_10000	1.046	0.912	0.778	0.885	1.069	0.812*	0.829*	0.812*	0.996	0.971	1.484*	0.75
LR_10000	0.961	1.133	1.034	1.141*	1.211*	1.438*	1.252*	1.259*	1.123	1.204	0.843	1.283
Agr_5000	1.003	0.972*	1.009	0.992	1.023	0.972	0.988	1.004	0.985	0.996	0.983	0.909*
Ind/Com_1000	0.951*	0.998	1.001	0.999	1.006	1.016	0.995	0.973*	0.988	1.04*	0.964	0.975
Highdr_1000	0.991	1.002	1	1.029*	0.997	1.074*	1.034*	1.014	0.963	1.027	0.955	1.034
TEMP_sum	1.621*	1.333*	1.509*	0.917	1.536*	0.738	0.89	0.694*	0.885	3.382*	1.527*	1.177
TP_win	4.645*	0.274*	0.656	0.557	1.116	0.125*	0.789	5.015	0.69	0.42	4.882*	0.069*
Topex	1.005	1.002	0.985*	0.989	0.973*	0.999	0.989	0.997	0.979*	0.994	1.01	1.005
Altitude	1.005*	1.005	1.002*	0.998	0.999	1.004*	1	0.989*	0.994*	1.008*	1.008*	0.987*
WS_win	2.709*	1.832	1.695*	0.094*	0.731	1.346	1.128	0.735	0.065*	1.109	2.379*	4.08E-05
SR_100	0.995	1.004	1.026	1.051*	1.025	1.046*	0.983	1.015	0.972	0.944	1.08*	0.982
Lowdr_1000	0.966*	0.993	0.985*	1.016*	0.986	1.01	0.987*	0.994	0.999	0.996	0.946*	1.031
Herb_5000	1.023	0.971*	1.001	0.963*	0.987	1	1.004	0.983	0.91*	0.931*	1.014	0.886*
MR_10000	1.031	0.847	0.987	1.294*	1.062	1.149	1.059	1.129	1.223	1.054	0.668	1.127
SLR_10000	1.458*	0.894	1.1	1.422*	1.155	1.364*	1.425*	1.46*	1.124	1.076	0.596*	1.431
Lowdr_10000	1.458	0.983	0.984	0.947	0.949	0.922*	0.946*	0.994	0.934	1.043	1.042	0.807*
Forest_10000	1.458	0.964*	1.01	0.991	1.054*	0.98	1.012	1.042*	1.018	0.956	1.002	0.96
FOREST1000	1.458*	1.025*	1.035*	1.026	1.031*	1.01	0.986	1.015	0.997	1.021	1.032*	1.05

*Wald statistic is significant at the 0.05 level.

Table 4.15 Confusion matrix for site type classification (% of MLOR site types by HCA site type)

Site type HCA	1	2	3	4	5	6	7	8	9	10	11	12	13	the main areas of confusion
MLOR														
1	**51**	5	13	3	10			2		2		15		1,3,12
2	2	**46**	17	4	8	2	2	2	4	4	2	7	2	3
3	6	15	**34**	2	26	3		1	4	4	3	3		2,5
4		7	3	**63**	4	2	4	3	2	10	2	1		
5	1	4	11	9	**56**	1		4	10	1		3		3,9
6	2	12	26	5	9	**26**	2	3	5	7	2	2		10
7		3		3	6	2	**49**	14	14		9			8,9
8	1	7	5	7	10	2	8	**33**	22		4	2		9
9	1	3	3	3	13		5	3	**68**	1		1		5
10		3	8	29	11	3		3	2	**29**	2		11	4
11		24		2		6	11				**58**			2,7
12	6	6	6		2			4				**75**		
13		2		2		2	2			2	2		**86**	

PART 2: MODELLING

5 Spatial model

This chapter describes the spatial model developed for the study region of Western Europe using a GIS-based approach (i.e. land use regression, 'LUR'). The spatial model developed here was developed at a fine resolution of 100*100m across Western Europe. This spatial resolution was used for three main reasons: 1) to give a highly localised estimation of ozone concentrations that might represent exposure at the semi-individual level; 2) to reflect the inter-urban variation in ozone concentrations, which depends primarily on local variations in concentrations of precursors and the factors (e.g. emission sources) which produce them; 3) 100m is considered to be the highest resolution achievable with the available input data, in which land cover data exerts the main limitation on accuracy (since this represents sources of ozone precursors).

Whilst using LUR to model air pollution is not novel, relatively few attempts have been made to apply it to ozone. The approach in this study is also unique in its application to O_3 at a fine spatial scale across a large study area, such as Western Europe. It is also the first study of this type to consider both traffic and background sites. The LUR model provides estimates of the long term mean O_3 concentration for years 2001-2007 and is constructed using data from 1211 monitoring sites, together with a range of land cover and other data.

5.1 Introduction

The results of the VCA in Chapter 4 suggested that marked spatial variations in O_3 concentrations occur across Europe. These variations also occur at different spatial scales, from the global (or hemispheric), reflecting broad-scale climatic patterns, to the local, reflecting proximal effects of emission sources. Such variations are also to be expected, because of the influence of geographic variations in the factors that determine O_3 production and loss – for example, emissions of O_3 precursors and scavengers, atmospheric chemistry and dispersion conditions.

Monitoring alone will never be adequate to represent this spatial variation and be able to provide the predictions of O_3 concentrations often required for policy assessment, to analyse pollution trends or for exposure assessment at small area or individual level for large epidemiological studies. As detailed in Section 2.2.2, some form of spatial modelling is therefore essential to extrapolate from these monitoring sites to unmonitored locations, and to map the spatial distribution of pollutants. This modelling needs to be able to deal with the inherent limitations of the available

data, as well as the resource and logistical constraints faced by most users (e.g. limited time, finances, expertise and computing facilities).

To create an O_3 map, both the spatial and temporal dimensions must be considered. This chapter focuses on the spatial dimension. There are three key aspects in terms of spatial dimension: the first relates to the extent of study area (geographic scale); second is the number of sampling points within the study area (measurements scale); and the final aspect is the optimal sampling interval between sampling points which reflects the complexity of spatial variation (operational scale) (Pualy and Drueke, 1996 cited in Diem, 2003).

For O_3, the operational scale is the most important because this needs to match the scale at which the processes controlling the formation and destruction of O_3 actually function, and thus the field lengths over which variation occurs. Estimates of this scale for O_3 vary. In a study in the city of Badajoz, Spain, Moral García et al. (2008) used a semivariogram to explore the spatial structure in 138 measurements taken using an automated portable analyser, and reported that the spatial range (i.e. the maximum distance over which significant spatial dependence in measured concentrations could be observed) was between 302 and 790 m. More generally, however, it may be expected that different operational scales can be defined, representing the influence of different determinants in different contexts. In urban areas, for example, O_3 precursors (especially NOx) can be expected to vary substantially over distances of a few tens to a few hundreds of metres, and O_3 concentrations can be expected to follow suit. In flat rural areas, on the other hand, concentrations are controlled by more regional factors, such as the meteorology and effects of inter-regional transport of O_3. In these situations, significant variation will tend to occur over field lengths of several (of even several tens of) kilometres. This suggests that modelling needs to be able to take account of these different scales, depending on context.

Diem (2003) conducted a review of 50 studies in which spatial O_3 models were developed by different approaches, including simple interpolation using IDW, kriging (from simple kriging to co-kriging), and multiple linear regression. He noted that, as a dense network of monitoring sites is not available in all study areas, and methods such as IDW and kriging are strongly affected by the distribution of sampling points, it may be difficult accurately to reflect the operational scale with these methods. All these approaches assume some degree of spatial autocorrelation (Griffith and Layne., 1999), but if this is over distances less than (or close to) the interval between monitoring sites, the resulting surfaces are liable to be subject to considerable error. To enable reliable estimation, therefore, the density of monitoring sites needs to vary, with higher densities in urban

areas, and perhaps in hilly or mountainous zones, where O_3 concentrations vary over shorter scales. For a large study area, such as the whole of Europe, monitoring networks will rarely be sufficiently well designed or developed to achieve this. Methods such as IDW or simple kriging, which rely solely on the monitored data, without making use of additional information in covariates, are thus likely to be unreliable. Thus an alternative approach in which the operational scales are better represented is required. A method using multiple linear regression is the approach of choice in modelling the spatial distribution of O_3 in this context (Diem and Comrie, 2002).

One of the most promising approaches to develop a spatial O_3 model is through the use of GIS technology, and within this one of the most widely applied methods in recent years has been LUR. Indeed, in several studies, LUR has been found to predict measured pollutant concentrations better than dispersion models (Gulliver et al., 2011, Cyrys et al., 2005). In many situations LUR is also often an attractive and appropriate technique because it is far less demanding in terms of data and computation than many other techniques, and can better deal with sparely or un evenly distributed monitoring sites.

It is helpful to recognise three main components of variability in the spatial dimension: trend (or drift), random spatially correlated variation, and noise (Burrough and McDonnel, 1998). Trend refers to the systematic variation over relatively large (e.g. regional or greater) spatial scales; for example, the trend for O_3 could occur from north to south across Europe, in response to broad climatic differences. The random spatially correlated component is the variation which is random but shows some degree of more localised predictability based upon relevant spatial covariates. Examples might include weather and/or emission related variables, representing the physical and chemical processes in O_3 formation, transportation and dispersion. The noise component is random and not predictable through LUR or other statistical techniques.

To build an effective spatial model, accounting for these sources of variability, attention has to be focused on the input data (i.e. predictor variables). LUR models have mainly been used to model primary pollutants such as NO_x or fine particles (PM_{10} or $PM_{2.5}$) of secondary pollutants such as NO_2 that form quickly and in close spatial association with emission sources. In these cases, the models typically use potential predictors related mainly to emission sources, sometimes along with variables representing the physical geography and meteorology (Hoek et al., 2008, Jerrett et al., 2004). The key variables, which typically explain most of the variation in the pollutant concentrations, thus comprise indicators for traffic such as road length, road intensity, distance to road or traffic volume; more general indicators of human activities (e.g. population or housing density); or land use data

(e.g. urban, industrial, open space, industry, commercial) (Hoek et al., 2008). Typically, also, these variables are measured at a local scale, since their influence is very localised.

Like NO_2, O_3 is also a secondary pollutant. The chemistry of O_3, however, is more complex as it is produced from the reaction of precursors, including NOx and VOC, in the presence of sunlight, as outlined in Section 2.1.1. Many of the influences are also more regional in scale. The predictor variables used in modelling O_3 therefore have to include information about the distribution of sources and/or emission rates of relevant precursors, as well as meteorological and other factors (e.g. topography) that influence reaction rates and dispersion, often at a broader scale.

Diem and Comrie (2002), for example, modelled maximum O_3 concentrations across a 500m resolution grid in Tucson, USA between April and August 1995-1998 during afternoon periods, using exogenous variables. O_3 concentration was obtained from seven urban sites surrounding by agricultural and mining land and ranging in elevation from 600 to 2800 metres above sea level. The estimated emission data for O_3 precursors were produced from the total country-wide emission for 1995 and from other regional inventories. These were spatially disaggregated to provide region-wide emissions totals for area sources, and then distributed to various point sources within cells (see Diem and Comrie (2001) for more detail). Meteorological data, population estimates, altitude and land cover data, as well as road length, were also used as proxies for O_3 transport and exposure. In addition, the cell exposed directly to air pollution was defined by using the altitude difference (i.e. topex).

An LUR model was then created by first clustering the days based on emission/meteorological data into five clusters, with each cluster representing different months. Secondly, 200 predictor variables, developed from the factors outlined above, were reduced to a smaller number of uncorrelated components, using principle components analysis. A model was then developed for each cluster of days using the deviation from average daily maximum O_3 concentration. At the end, the modelled O_3 concentrations were added back to the average maximum concentrations to produce the final predicted concentration. Each cluster-specific model involved 5-10 predictors mostly proxy variables (from land cover data and road length) and variables relating to the short-long distance transport of O_3 precursors from their emission source (based on meteorological data). The overall coefficient of determination R^2 was 0.9, and the RMSE was ~9 $\mu g/m^3$.

As identified by Diem (2003), the problem with using these types of exogenous variables in modelling O_3 concentration is often their coarse resolution. While this may help to pick up their

regional effect, it may mask more local influences, and thus smooth out much of the local variability in O_3 concentrations. Some type of interpolation is therefore needed to rescale the variables to match the modelling resolution. Emissions data are also often known to be prone to errors (Diem, 2003). Therefore, if using proxies for emissions as independent variables, attempts should be made to use data that are well correlated with O_3 concentrations and have relatively low inherent error (e.g. fine spatial resolution).

To date only one other study has used LUR to develop a spatial O_3 model in coras resolution (1*1 Km). Like the present study, this, too, was for the whole of Europe (Beelen et al., 2009), and it also echoed many of the principles put forward by Diem (2003). Predictions were derived by bringing together three separate LUR models, one representing the broad regional pattern of variations (a so-called global model), calibrated using data for remote rural sites; another representing more local variation in rural areas; and a third representing variation in urban areas. Both the predictors and buffer sizes were selected to reflect these different scales and contexts. Meteorological variables, altitude and distance to sea were used to model the global variation; land cover and transport variables, for relatively broad buffer zones (5 and 21 km diameter), were used to model additional variation, over and above the global pattern, in rural areas; and the same variables at higher spatial resolution (1km buffer zones) were used to model variation in urban areas. Global and rural models were calibrated using only rural background sites. The urban model was developed using only urban background sites. Models were also built under the constraint that variables entering the regression equation had to conform with a priori defined directions of effect, The most significant predictors in explaining the variability in O_3 were found to be altitude, distance to sea, major roads, high density residential land, and variables reflecting particular weather regimes. Their work thus demonstrated that LUR is a promising approach in modelling O_3, and it is used as the starting point for modelling here.

5.2 Methodology

The LUR methodology, used to derive estimates of the long-term average O_3 concentration of shown in Figure 5.1.

Multiple linear regression analysis is one of the most widely used methodologies in modelling a dependent (predicted) variable on the basis of several independent (predictors) variables. This

statistical method assumes that the relationship between the dependent and each independent variable is linear.

Using regression analysis to model O_3 requires the construction of an equation which can predict variation in O_3 concentrations on the basis of the predictor variables. Multiple regression models for the effects of k predictor variables take the general form as follows:

$$Y = ß_o + ß_1 x_1 + ß_2 x_2 + \ldots\ldots\ldots\ldots\ldots + ß_k x_k \qquad \textbf{Equation 5-1}$$

where:

Y	is the outcome (estimated long-term mean of O_3 concentration)
$ß_0$	is the intercept (the value of Y where x=0)
$ß_1$ $ß_2$...$ß_k$	are regression coefficients of the linear regression equation which explains the increase in Y for every increase unit in x
x_1 x_2...x_k	are the environmental (predictors) variables

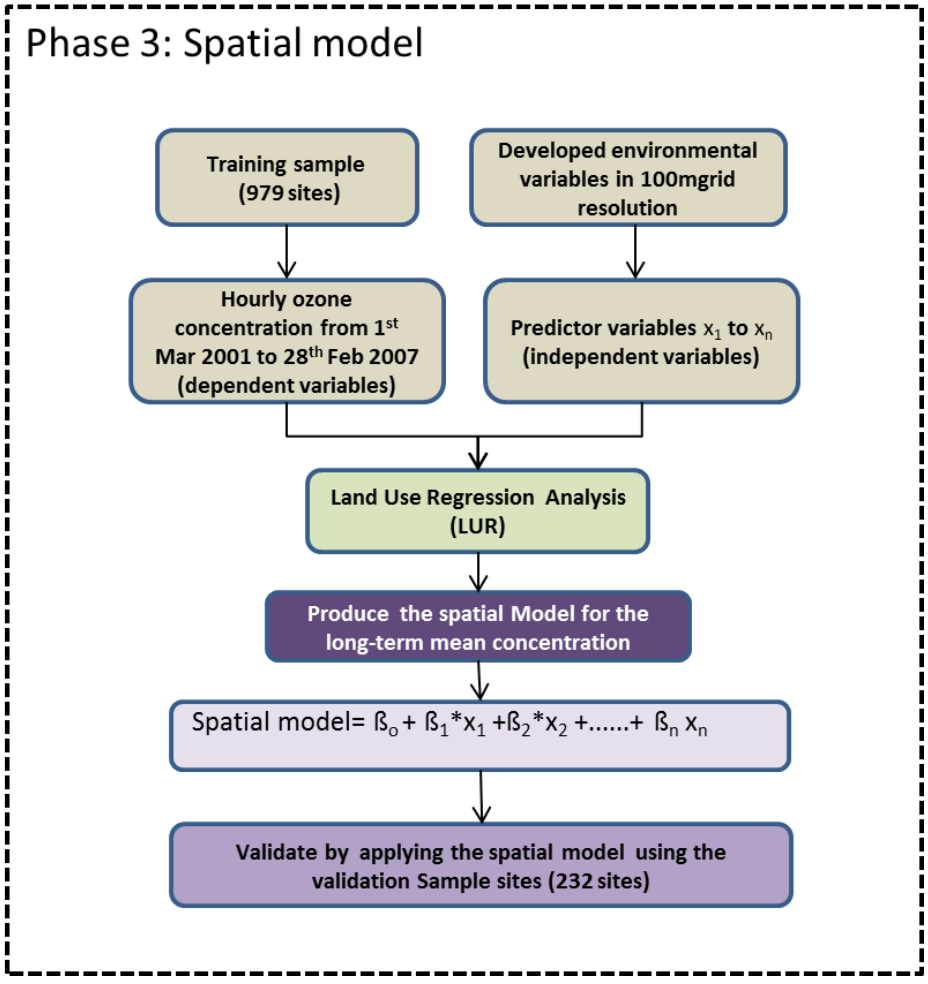


Figure 5.1 Methodology steps for spatial model in Phase 3

Because of the possibilities of bias in the model, it is important to test the performance of the model by validating predictions against an independent set of data. There are two main ways of doing this. One is by a process of cross validation. In this, a series of LUR models is built, using N-k sites – i.e. by dropping a specified number (k) of sites on each occasion, and their value being predicted from the model. The process is repeated N/k times, until each site has been dropped once. Comparison of the predictions with the observed values then provides an estimate of model performance. The alternative approach is to split the data into a training and validation set, using the first only for model building and the second only for testing of model performance.

Each approach has advantages and disadvantages. Cross validation has the advantage of using all the data both for model building and for validation, and therefore maximising the amount of information gained from the data. A difficulty arises, however, in specifying the final model, since the models built on each occasion are likely to vary slightly, not only in terms of the coefficients attached to each variable, but also the variables that were selected by the regression analysis. It is also more time-consuming, because model-building has to be done multiple times. The split-data approach clearly avoids these difficulties, but may result in less robust models and measures of performance, because the number of sites used in each case is smaller. Biases in the splitting of the sites into the training and validation data sets will also feed through into the models, and may create differences in model performance between the training and validation data sets.

On this case, the split-sample approach was selected for three reasons. The first is that this approach provides a single, unequivocal model, that can then be validated and used for prediction. The second is that the data set is large, so the problems of sample numbers are unlikely to be severe. The third is that similar validation was needed for other parts of the modelling – i.e. the time functions and models including meteorological data. It was considered helpful to build and test models for each of these on constant data sets, so that the performance statistics could be directly compared, and the improvements in predictions by incorporation of additional sub-models readily assessed. To enable validation of this model and all other modelling in the coming chapters, the 1211 O_3 monitoring sites were divided into a 979 sites as training data set (80% of sites) and a 232 sites as a validation set (20%). For more details see B, Section IV

5.2.1 Variables and data sources

Data on the annual average O_3 concentrations from the 979 training sites were used as the dependent variable in model-building. Some studies use a transformation (usually logarithm) of concentrations in an attempt to better approximate linearity of the relationship, to achieve a normal distribution of the residuals, and also to avoid negative predictions. Untransformed long term mean O_3 concentrations were used here, as there was no violation of the criterion for normality of the distribution. An additional advantage of using untransformed concentration data, as opposed to transformed concentrations, is that it is easier to interpret the relationship between concentration and the predictor variables because the relationship is directly additive, rather than proportional, as with logarithmic transformations. This is supported by theoretical arguments about how O_3 production occurs.

The environmental variables offered as predictors in the LUR were described in Section 3.2.3. These data consist of land cover, road length by type, meteorology, altitude, topographical exposure (topex) and distance to sea, measured for different window zones around each of the monitoring sites. Table 5.1 includes a description of each predictor, window sizes, and the required direction of effect in the final regression model. As already mentioned meteorological conditions often exert the major impact on concentrations. Relevant meteorological factors include solar radiation, temperature, precipitation and wind speed, precipitation (Section 3.2.3.5). Temperature and solar radiation are important factors in determining rates of photochemical reaction, which increases as solar radiation intensifies and as air temperature rises. Precipitation plays a role in wet dispersion and also acts a proxy of cloud cover, both of which reduce O_3 concentrations. Wind speed, on the other hand, is more complex. While it has often been shown to have a negative association with O_3 concentrations, largely perhaps because it encourages dispersion and dilution, in some cases its effect is positive. This mainly appears to be because it brings in O_3 from other, enrich areas or by promoting turbulence and the vertical mixing of O_3 from higher layers in the atmosphere.

Topography is also an important determinant, both in its own right and through associations with meteorology and human activities. Higher concentrations of O_3 tend to occur at higher altitudes (Coyle et al., 2002), largely because of the more intense solar radiation, but also because mountainous areas tend to contain few emission sources of NO or other scavengers. Topex, or topographic exposure, defines the degree of openness of the terrain, in terms of the relative relief. It can be expected to have a positive association with O_3 concentrations both because it implies locations that are higher than the surrounding land, and thus have higher levels of solar radiation,

and because these areas are less likely to contain major emission sources. Areas of negative topex (i.e. depressions or valleys), on the other hand, will be more shaded and thus have reduced solar radiation, and potentially act as transport corridors and the focus for settlement, thereby increasing emissions of NO and other scavengers. Distance to sea is a proxy both for the likelihood of influx of O_3 from the ocean (i.e. open sea), where O_3 formation is enhanced (Caballero et al., 2007), and for the influence of maritime climatic conditions.

Traffic is the major source of O_3 precursors, especially NOx. It could be represented in a range of variables used in LUR such as traffic intensity or counts on the roads. Due to the difficulty in getting these data for the whole of Europe, and following the example of many LUR studies, road length by type was used as a proxy for traffic volume. Several studies have shown that road length is a good substitute for traffic, giving similar results in LUR models (Vienneau et al., 2010, Henderson et al., 2007, Madsen et al., 2007).

Finally, land cover data are used to account for other sources of O_3 precursors, especially from anthropogenic and biogenic sources. For example, high density and low density residential lands are proxies for emission from domestic activities (e.g. heating). Industrial, commercial, construction, and port areas, were combined together in one variable, termed industrial/commercial land to provide a proxy for industrial emissions. Forest and agricultural lands can also produce O_3 precursors, especially VOC. However, their effects may be more complex. In the growing season, for example, some trees and agricultural crops may also act to diminish O_3 concentrations by encouraging deposition by absorbing O_3 through the stomata of vegetation (Coyle et. al., 2002).

Some LUR studies for other pollutants also used population as a predictor variable (Skene et al., 2010). If the LUR model is to be used in an epidemiological study with an ecological design, or in a health impact assessment for the whole population, the use of population as a predictor can cause difficulties, for it results in a degree of duplication. It was thus decided a priori that population would not be used in this LUR model for O_3. In any event, good land cover data will usually provide an adequate and higher resolution measure of population distribution.

All of the above mentioned data sets are available across the whole study area. The regression equation thus developed from the LUR model can therefore be used to predict O_3 concentrations at unmonitored locations (i.e. all 100 metre grid cells) across Western Europe for mapping purposes. A summary of the available data is provided in Table 5.1.

5.2.2 Model building

The LUR procedure used here follows that by Hoek et al. (2011), which is also used in the ESCAPE[23] project. ESCAPE is a European Study of Cohorts for Air Pollution Effects, focusing primarily on fine particles, particle composition, and nitrogen oxides. As the O_3 estimates deriving from this work were also to be used in health studies within the ESCAPE project, it was considered important to follow the ESCAPE LUR methodology. The steps described below are largely a reproduction of the procedure as set out in the ESCAPE exposure manual.

As already indicated, measurements of the average concentrations of O_3 (long-term mean from March 2001 to February 2007) from a set of monitoring sites were used to train the LUR model, and the model then validated against data from an independent set of sites. A supervised, stepwise approach was used. First the bivariate correlation between observed O_3 concentration and each possible predictor variable was performed and the adjusted R^2 recorded. The predictor giving the highest correlation with observed O_3 concentration was then selected for entry into a multiple regression analysis. Additional predictors were then added, one at a time, in subsequent steps on the basis of a previously defined set of selection criteria. The process was repeated until there were no remaining predictors that met the selection criteria, namely:

- Each predictor has to show a significant marginal correlation with the observed concentrations ($P \leq 0.05$);
- The direction of the effect must be as expected according to a priori considerations (Table 5.1);
- The significance and direction of effect for the predictors already in the model should not change when new predictors are included at each step;
- The inclusion of the variable into the model must increase the adjusted R^2 by at least 0.01.

Because variables are available for many windows, there is the possibility that the same predictor variable may be included in the model at different window distances. If this happens, the difference between the larger window and small window is calculated. For example, if high density residential land within a 5000m (Highdr_5000) and a 300m (Highdr_300) window are included, in the final model the Highdr_5000 will be replaced with the 5000m minus 300m window (e.g. Highdr5000-300). However, it has been shown that using complete (nested windows) or disjoint windows (rings

[23] http://www.escapeproject.eu/manuals/ESCAPE_Exposure-manualv9.pdf

without gaps) in regression models provides the same results (von Klot, 2011). The advantage of the approach used here is that interpretation can be easier, for example in estimating the contribution to O_3 concentrations from different window zones.

Table 5.1 Predictor variables used in LUR

	Abbreviation	Predictor variable	Window's radius sizes(m)	Direction of effect
Land cover variables[a]	Highdr	High density residential land	**100,300,500,1000,5000,10000**	Negative
	Lowdr	Low density residential land		Negative
	Ind/Com	Industrial and commercial land		Negative or positive
	Herb	Herbaceous land		Negative or positive
	Agri	Agricultural land		Negative or positive
	Forst	Forest land		Positive
	Opsp	Open Space		-
Topographical variables	D2S	Distance to sea	**Kilometre**	Positive
	Alt	Altitude (height above sea level)	**metre**	Positive
	Topex	topex (topographic exposure)	**metre**	Positive
Road length variables[b]	MR	Motorways	**100,300,500,1000,5000,10000**	Negative
	SR	Secondary roads		Negative
	LR	Local Roads		Negative
Meteorological factor[c]	SSR	Surface solar radiation	**W/s**	Positive
	TP	Total precipitation	**mm**	Negative
	TMP	Temperature	**C°**	Positive
	WS	Wind speed	**m/s**	Negative or positive

[a.] in Percentage
[c.] in metre
[b.] All meteorological factors were calculated for three averaging periods: annual (_ann), summer (_sum), and winter(_win), as mentioned in section 3.2.3.5

By using a minimum value of change in adjusted R^2 as an inclusion criterion, some predictors may become non-significant as other variables are included in the model. The final step was thus to evaluate the significance of all variables in the model. Variables with $P > 0.05$ were sequentially removed from the model, starting with the least significant, until all predictor variables in the model had a $P \leq 0.05$.

As noted in Section 3.2.3.1, land cover data were obtained from the CORINE data base. In this, residential urban areas are described by two classes. High density residential areas are classified as continuous urban fabric; low density residential areas comprise the discontinuous urban fabric. Some discrepancies in the definition of these two classes seem to exist across the EU. In the Netherlands, for example, only discontinuous urban fabric is recognised (Vienneau, et al., 2010). This could be because the structure of residential land there is, in fact, different from other countries in Europe. As each country submits their own land cover data to Europe for consolidation into CORINE, however, it could also be due to differences in interpretation of the classification methodology. To overcome this problem it was decided, a priori, that the low density residential predictor would be forced to enter in the case of the occurrence of high density residential in the final model, at the end of supervised stepwise inclusion of significant predictors.

Once the LUR model had been built, predicted O_3 concentrations were mapped by applying the final LUR equation to the relevant predictor grids, using ArcMap grid arithmetic commands. The estimated O_3 concentrations for 100m grid cells over the whole study area were thus produced.

The model was evaluated by comparing predicted concentrations against the observed concentrations for the reserved validation dataset. The full list of validation metrics suggested by Fox (1981) and Willmott (1982),and widely used for model testing, were calculated (Table 5.2).

Linear regression assumes independence of the residuals; thus normality of the residuals was assessed by obtaining the histogram. Moran's I for spatial autocorrelation was also used to check whether the residuals exhibited spatial autocorrelation (Jerrett et al., 2007, Ross et al., 2005). Moran's I is a statistical indicator with a range between -1 and +1, where 0 means no correlation with nearby sampled locations (i.e. the residuals are not spatially auto-correlated when Moran's I is nil).

Table 5.2 Model performance quantitative metrics

Metric	explanation	Equation	Purpose	Range
Summary measures[a]	Observed conc. mean ($\bar{O}$) and predicted ($\bar{P}$) mean	$x\,mean = \frac{1}{N}\sum_{i=1}^{N} x_i$	To measure centre tendency	No range
SD	Standard deviation of observed (SD_o) and predicted (SD_p) conc.	$SD = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - x\,mean)^2$	To measure the variation	No range
R^2	Adjusted regression coefficient squared	SPSS outputs	To measure the explanation variability explained by model	No range
RMSE[b]	Root mean square error	$[N^{-1}\sum_{i=1}^{N}(P_i - O_i)^2]^{0.5}$	To measure the average error between predicted and observed variables and sensitive to extreme.	Take value from 0 to infinity
d[c]	Index of agreement	$1 - \left[\frac{\sum_{i=1}^{N}(\acute{P}_i - \acute{o}_i)^2}{\sum_{i=1}^{N}(\lvert P_i\rvert + \lvert O_i\rvert)^2}\right]$, $0 \le d \le 1$	a standardized measure of the degree of model prediction error	varies between 0 and 1 where: 1 indicates a perfect match, and 0 indicates no agreement
VIF	Variance inflation factors	SPSS outputs	Measure the multicolinarity	<5

[a] N the number of cases and x_i is predicted or observed value for i case
[b] P_i predicted value for i case and O_i observed value for i case
[c] $P'_i = P_i - \bar{P}$ and $O'_i = O_i - \bar{O}$

5.3 Results

The final prediction model explained 67% of the observed variability in O_3 concentrations (adjusted R^2=0.67) with RMSE=7.59 μg/m^3.

Table 5.3 provides summary statistics for the final model. The VIFs for each variable are below 2, indicating that no multicolinarity between the predictors was observed. The table also includes the standardised coefficient (Beta). This shows the impact of a one standard deviation change in the predictor variable on the long-term mean of O_3 concentrations. It can thus be used to compare the relative importance of the predictors, as shown in Figure 5.2. On this basis, it is apparent that spatial variations in O_3 concentration are characterised mainly by altitude, followed by local road density (which is a proxy of NOx emission), and summer temperature.

Table 5.3 Summary of LUR model

Predictor	Unstandardised Coefficients β	Std. Error	Standardised Coefficients (Beta)	P-value (Sig)	Sub sequential Adj.R^2	Collinearity Statistics (VIF)
Constant	18.17	3.03		0.00		
Alt	0.02	0.00	0.50	0.00	0.28	1.91
LR_10000m	-2.88E-05	0.00	-0.37	0.00	0.20	1.54
Topex	0.04	0.01	0.15	0.00	0.04	1.17
TMP_sum	1.54	0.11	0.31	0.00	0.04	1.43
WS_ann	2.79	0.35	0.19	0.00	0.03	1.64
Highdr_1000m	-0.15	0.02	-0.22	0.00	0.03	1.60
Lowdr_1000m	-0.02	0.01	-0.03	0.10	0.01	1.62
Dis2sea	-0.01	0.00	-0.12	0.00	0.01	1.36
Forest_1000m	0.02	0.01	0.09	0.00	0.01	1.82
IND/COM_5000m	-0.23	0.05	-0.12	0.00	0.01	1.58
Agri_5000m	-0.05	0.01	-0.09	0.00	0.01	1.51

Dependent variable: long-term mean of O_3 concentration from March 2001 to February 2007

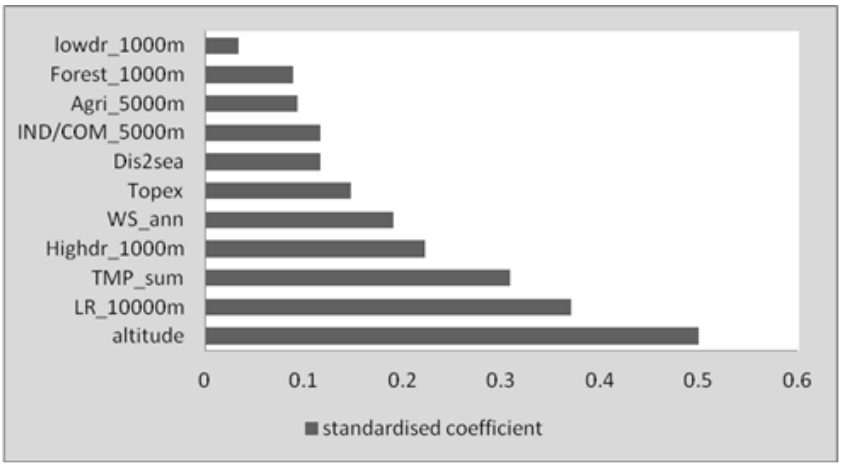


Figure 5.2 Importance of predictors in long term O_3 LUR model

The unstandardized coefficients (β) define the slope of the regression curve, and can be used to show how concentrations change with a one unit change in the predictor. An increase of one kilometre in altitude, for example, increases the long-term concentrations of O_3 by 0.02 * 1000 = 20 µg/m³. An increase in topex by 50 metre (i.e. as sites become more open and exposed) increases the concentrations by 0.04*50= 2 µg/m³. Both variables probably reflect the general tendency for O_3 concentrations to increase under conditions of high solar radiation, and in more remote or exposed areas where concentrations of other pollutants are reduced. In contrast, a one kilometre increase in road length within a window radius of 10Km causes an 0.03µg/m³ decrease in O_3 concentration. This is probably due to the fact that road length is a proxy of NO emissions, which results in scavenging of O_3; 90% of NOx emissions derive from transport in the form of NO (Vestreng et al., 2008). It is also

notable that, only local roads were included in the model and major roads were not significant predictors. This was presumably due to a number of inter-related reasons. The first is that major roads (as defined in the data used here) are relatively scarce, so around the large majority of sites are either absent or occur only as short distances; local roads, on the other hand, are far more ubiquitous, and probably give a better, general indication of emissions from road sources. A second reason is that traffic flow on local roads is slower than on major roads, and are thus likely to produce more emissions of O_3 scavenging precursors (e.g. NO) per unit of traffic flow than do major roads. Thirdly, it may indicate that, in modelling O_3, traffic intensity is more important than the road length; in the case of local roads, a more direct relationship between length and traffic volume probably occurs. The fourth reason is that local road density is especially variable in urban areas, where these roads act as important traffic conduits: the density of roads is thus an especially good indicator of traffic density in urban areas, in comparison to major roads which typically reflect only inter-urban traffic flows.

Distance to the sea (Dis2sea) also enters the model, with a negative sign indicating that O_3 concentrations decrease with increasing distance from the open sea, as expected. A 10 kilometre increase of distance to sea reduces the long term concentration by 0.1 μg/m³. The land use variables all tend to reduce O_3 concentrations, with the exception of forest land. An increase by 10 percent of forest area within a window of 1Km radius (Forest_1000m) tends to increase O_3 concentrations by 0.2μg/m³. This is probably for two reasons: most importantly, perhaps, forest areas represent areas with little or no local emissions, so scavenging is limited and O_3 concentrations raised; secondly, trees (e.g. broad-leaved forest and coniferous forest) can be important sources of biogenic VOC (e.g. isoprene and monoterpenes) emissions, which may encourage O_3 formation. On the other hand, a 10 % increase in the area of agricultural land within a 5Km window radius (Agr_5000m) tends to decrease O_3 concentrations by 0.5μg/m³. This is possibly because small plants play an important role in dry deposition.

An increase in the area of industrial and commercial land (as a proxy for NO and VOC emissions) by 10 percent within a window of 5Km radius, decreases long-term mean O_3 concentrations by 2.3 μg/m³. In addition, every 10 percent increase in high density residential land (Highdr_1000m) and low density residential land (Lowdr_1000m) within a 1Km window radius tends to reduce O_3 concentrations by 1.5 and 0.2 μg/m³, respectively. This probably reflects the fact that built up land is a source of NOx emissions. Also, Highdr has a greater tendency to decrease O_3 concentrations compared to Lowdr, as is to be expected given the different levels of emission implied.

Meteorological variables have a significant role in increasing O_3 concentrations. For every one Celsius increase in summer temperature and one metre per second increase in wind speed, O_3 concentrations are increased by 1.54 and 2.79 µg/m³, respectively.

O_3 concentrations derived from the final LUR model are mapped in Figure 5.3. This figure shows that O_3 concentrations across Europe tend to increase towards the south-east, due to generally higher temperatures in south-east Europe. Moreover, the maritime effect is more obvious along the north-west coast (e.g. in Scotland and Ireland), perhaps because of the stronger influence of the Gulf Stream which acts to raise temperatures along the coast, and the effect of prevailing south-westerly winds that carry O_3 inland from North America (Wild and Akimoto, 2001 cited in Monks et al., 2009).

Stability of the model was evaluated by the application of the model to validation sites. The model was found to perform similarly to the training sites, explaining 65% of the observed variability in O_3 concentrations with RSME=7.7µg/m^3 (see Table 5.4). This suggests that the model is not over-fitted to the training data. According to the index of agreement (d) the match between predicted and observed concentrations of O_3 is perfect (=1) in both samples training and validation sites. The overall distribution of the residuals is shown in Figure 5.4. The standardised residuals exhibit in general a normal distribution with a mean concentration of 0 and a standard deviation of 1. Figure 5.5 shows the scatterplot of the observed against predicted O_3 concentrations at all of training monitoring sites. There are no notable outliers and the prediction quality is good. No spatial autocorrelation in the residuals was found, with a non-significant Moran's I = 0.2 (z-score of 0.65), indicating that the pattern does not appear to depart significantly from random. This emphasises that modelling O_3 with a finer spatial resolution helps to remove any significant spatial correlation in the residual compared to models using a coarser resolution 1*1Km (e.g. Beleen et al, 2009).

Table 5.4 Performance metrics for the spatial model (LUR)

Metrics	Training sample		Validation sample	
	Observed values	Model predicted values	Observed values	Model predicted values
Mean	49.6 µg/m³	49.6µg/m³	49.0 µg/m³	50.4 µg/m³
StD	13.1 µg/m³	10.7 µg/m³	12.1 µg/m³	10.7 µg/m³
Adjusted R^2	0.67		0.65	
RMSE	7.5 µg/m³		7.7 µg/m³	
d^2	1		1	

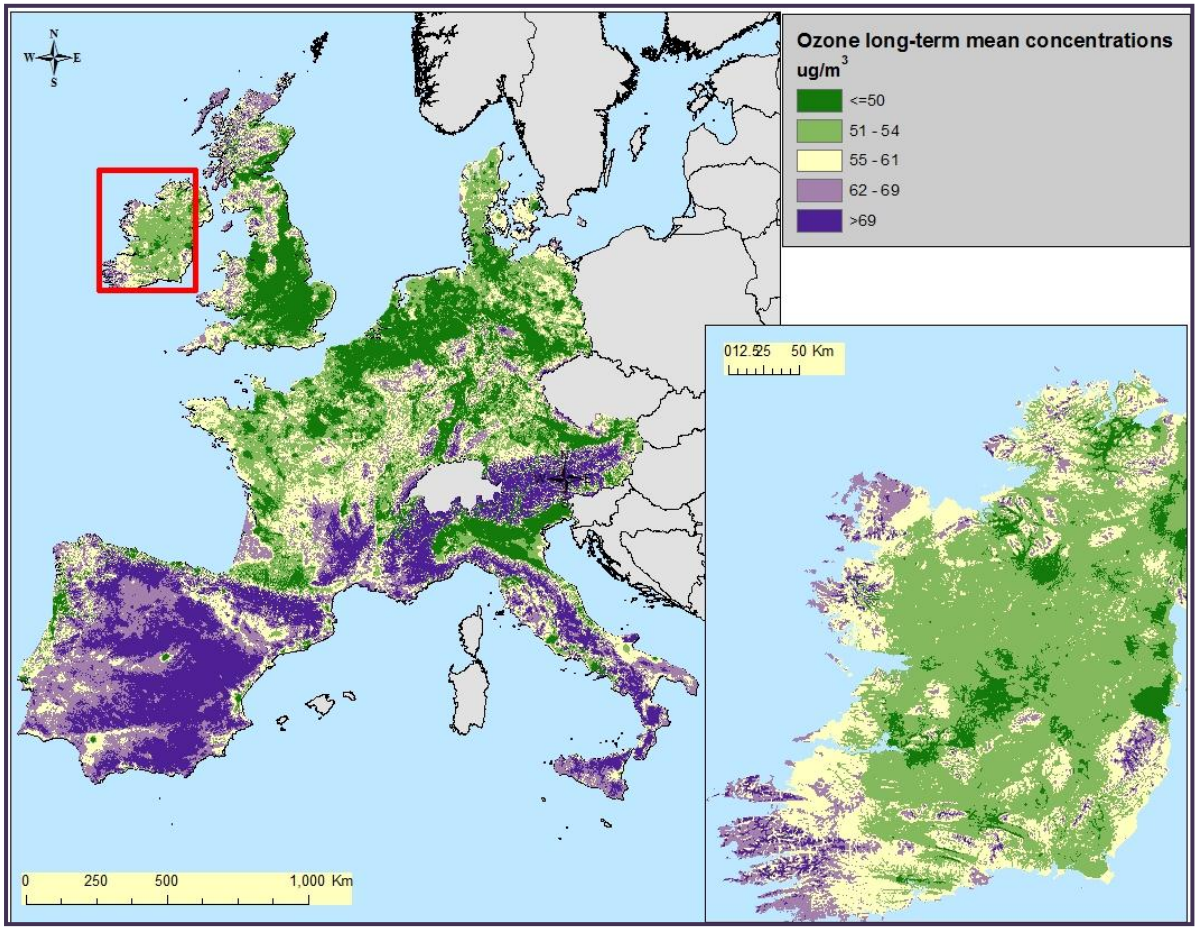


Figure 5.3 EU map of modelled long-term O_3 concentrations with 100m grid resolution

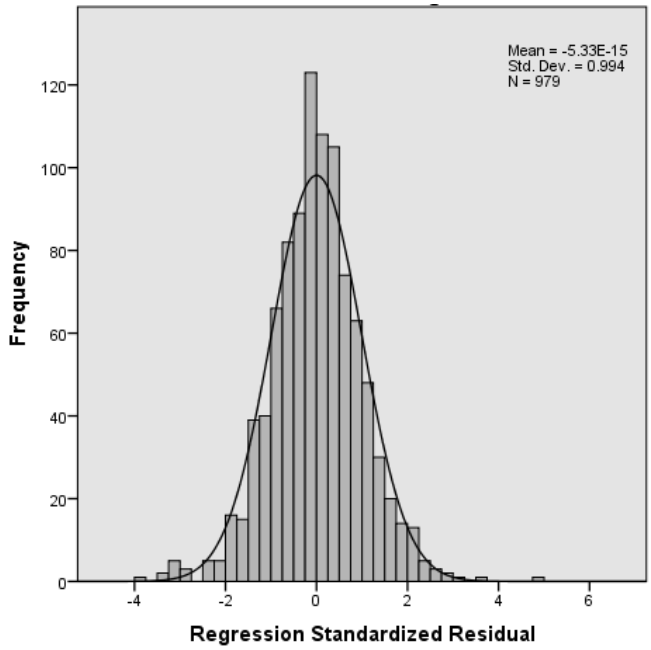


Figure 5.4 Histogram of standardized residuals (For training dataset)

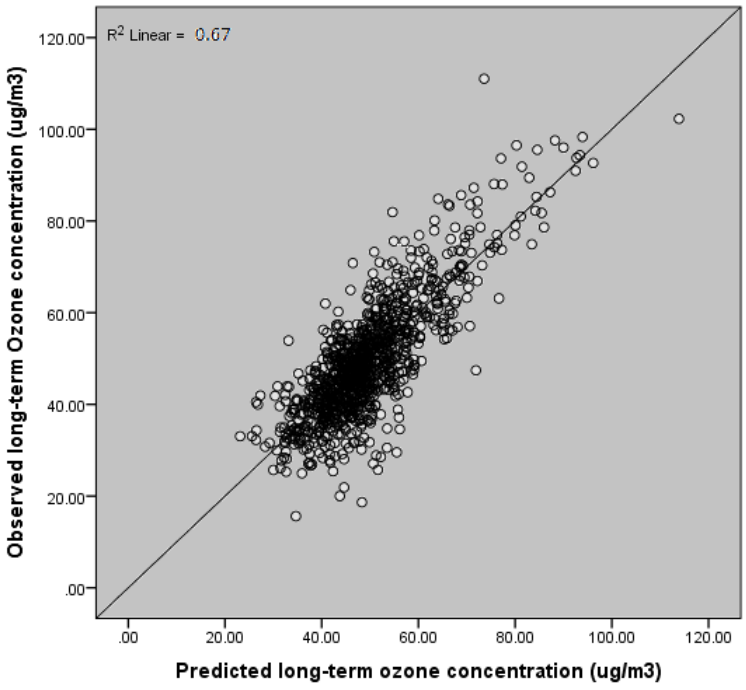


Figure 5.5 Scatter plots between observed and predicted concentrations

5.4 Sensitivity analyses for the global LUR

5.4.1 Components of variability

As mentioned, the underlying principle of spatial modelling is to attempt to model the three types of variability: global trend, random spatially correlated variation, and noise. These three different sources of variability are important in the spatial dimension. The LUR model in the present study mainly describes the second of these: the random spatially correlated variation: there are no terms for spatial trend (e.g. based on latitude or longitude) and the noise cannot be predicted. An indirect indication of spatial trend may, however, be given by the variables for solar radiation and temperature, both of which show a general increase from north to south in Europe. However, they also pick up a more complex trend, given that both temperature and solar radiation vary at different scales: locally (in response to local topography), regionally (in response to broader topographic and atmospheric circulation patterns) and at the broad scale (in response to continental-scale effects such as global circulation and climate patterns. To explore whether any global trend remained after including these variables in the model, the residuals from the spatial model were therefore regressed against latitude and longitude. This resulted in an adjusted R^2 equal to 0.013. This means that including latitude and longitude would add a little over 1% to R^2 – enough under the criteria used here for it to be included in the model. This demonstrates that meteorology factors explain some of the global trend in these data.

5.4.2 Global versus local models

A single LUR model was developed here for the whole of the study area, and representing all site types, and as Figure 5.3 shows this provides a plausible map of O_3 concentrations without any marked disjunctions or anomalies. A single model for the whole of Europe of this type is likely to be appropriate for many applications – especially where consistency is the over-riding criterion: for example, for continental-scale risk assessments or large epidemiological studies. For many applications, however, interest may focus on smaller areas – for example, on individual countries, or specific types of environment (urban, rural etc). The question thus arises whether this global model would still be appropriate, or whether it might lead to biases and uncertainties that could only be reduced by developing a use-specific model. This question was investigated in a series of post hoc analysis, where sites were stratified by different criteria and the correlations between observed

and predicted concentrations were explored in different site-types; climatic regimes (by latitude); physical terrain (by altitude); land use (rural vs. urban), and geographic location (by different countries).

5.4.2.1 Variations by site type

Site type was shown to be an important determinant of the temporal pattern of O_3 concentrations and is used as a basis for stratifying sites for modelling temporal variability. It might thus be expected to have significance in terms of the spatial distribution of O_3 concentrations.

Figure 5.7 shows the relationship between modelled and observed concentrations coded by site type and Figure 5.8 summarises these data more simply, by presenting the means of the modelled and observed concentrations for each site type. Both indicate that the performance of the model is broadly consistent across the different site types, with no obvious bias in the estimates, and with a strong correlation between the mean of the modelled and observed concentrations of the thirteen site types ($R^2 = 0.93$).

The distribution of sites, however, is not even either by site type, with the numbers further reduced as sites have to be divided into training and validation datasets. This may produce differences in the quality of the model in different areas of Europe which will inevitably be weighted towards site types that provide more training sites. The global model may therefore be sub-optimal in less well-represented areas.

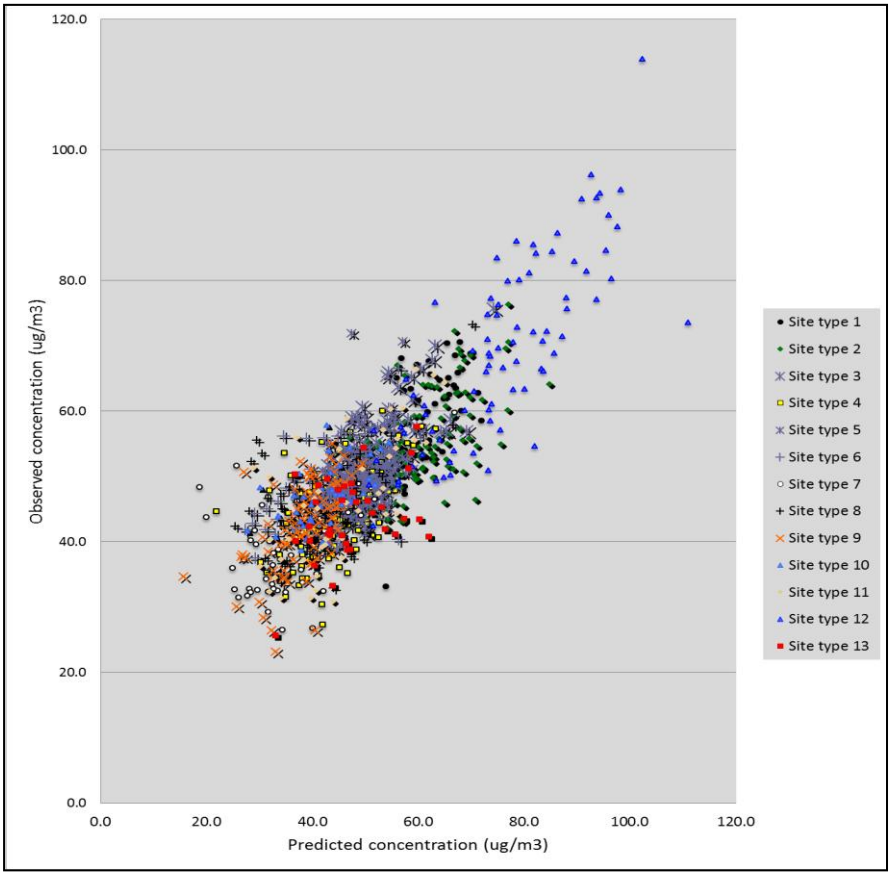


Figure 5.6 Scatterplot for observed against predicted long term concentrations coded by the thirteen site types

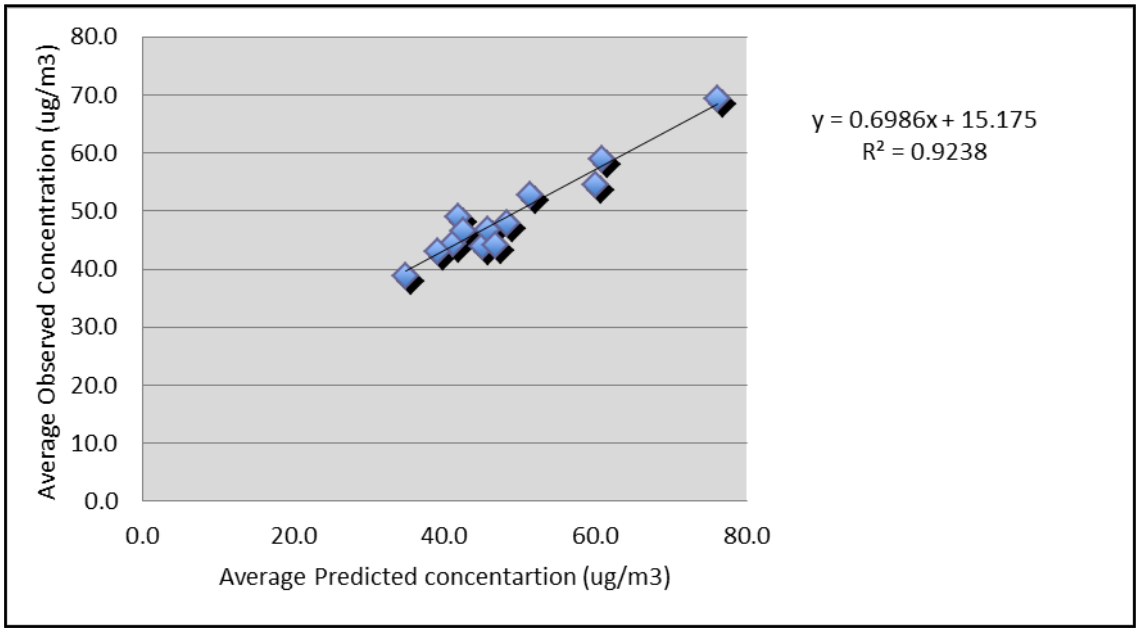


Figure 5.7 Averages of the observed and predicted concentrations for the thirteen site types

To investigate this issue further, therefore, boxplots of the residuals by site type were drawn (Figure 5.9). As can be seen, for most site types, residuals are broadly similar, with the large majority of residuals between -10 –and +10 μg/m^3.

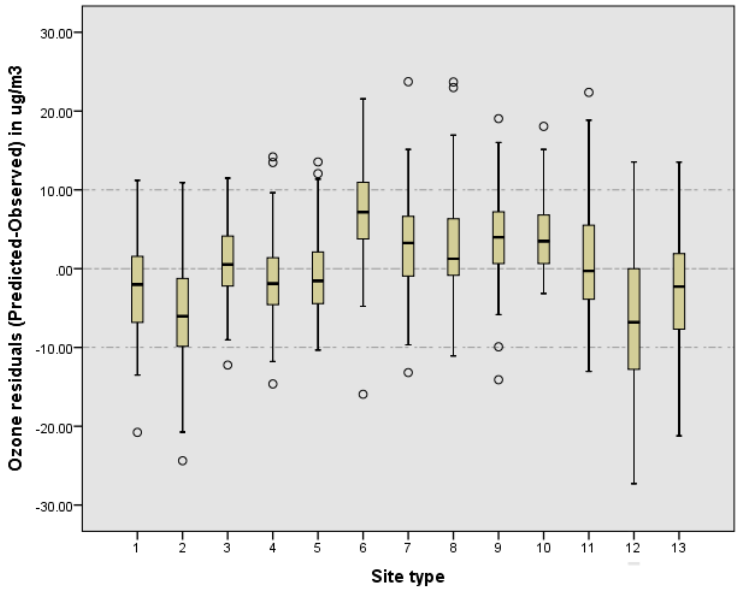


Figure 5.8 Boxplot of residual by site type

There is, however, a tendency to over-estimate concentrations in some site types, notably 6 and to a lesser extent, 7, 9 and 10. All these are highly urbanised site types, with heavy traffic volumes, as indicated by the long road lengths (Table 4.10). In contrast, the model tends to under-estimate the concentrations in site types 2 and 12 which are described as sunny mixed land use and forested mountains. They are, therefore, both areas characterised by high levels of photochemical activity. The suggestion is thus that the model over-estimates in areas of O_3 depletion, characterised by abundant sources of NO and other scavengers, but under-estimates in areas of O_3 formation. Interestingly, the same finding has been observed in the results of dispersion modelling (Daniel and Denise, 2006).

5.4.2.2 Variation by country

Variations in the performance of the model between different countries might also be expected, not only because these comprise different types of environment, but also because there may be

differences in the quality and character of the available data, and because of the large differences observed in the number of monitoring sites.

Table 5.5 summarises the performance of the model by each country in the study area. Some differences in performance are evident. The coefficient of determination (R^2) ranges between 0.47 (Portugal) and 0.84 (Ireland), while the RMSE varies from 4.2 μg/m^3 (Netherlands) to 10.1 μg/m^3 (Italy).

Table 5.5 Performance of the global LUR within different country

country	Density of sites[a] within 1000Km2	Adj.R ^{2}	RMSE
AT	0.98	.73	9.24
BE	0.92	.70	4.88
DE	0.63	.75	5.02
DK	0.09	.77	7.28
ES	0.37	.69	8.93
FR	0.53	.67	5.97
GB	0.20	.53	8.01
IE	0.08	.84	4.98
IT	0.21	.52	10.10
NL	0.65	.50	4.20
PT	0.22	.47	7.65

a. Training dataset

There is no association between site density and RMSE (R=-0.17) or R^2 (R=0.14). This suggests that the variations are not a result of differences in site density, but other instead relates to other reasons, such as the specific location of the sites, the inherent complexity of the terrain and land use, and different meteorological conditions in different countries.

5.4.2.3 Variations by latitude

Latitude may be expected to affect O_3 concentrations, and model performance, through its influence on meteorological conditions, and factors such as day length. To investigate effects on model performance, sites were divided into northern and southern groups, on the basis of latitude. One third (those with the highest latitude) was classified as northern and one-third (with the lowest latitude) was classified as southern.

Table 5.6 summarises the performance statistics. In terms of R^2, it is evident that the LUR model performs equally well in both regions. The RMSE, however, is somewhat higher in the southern sites (RMSE=8.85 compared to 5.87ug/m^3). This reflects the tendency of both weather conditions and O_3 concentrations to vary more strongly in the south, resulting in an increase in the estimation error.

Table 5.6 Performance of the global LUR in Northern vs. Southern region

Sites location	No. of sites	Adj. R^2	RMSE
Southern	300	0.67	8.85
Northern	300	0.67	5.87

5.4.2.4 Variation by altitude

Sites were categorised according to altitude to differentiate between low-lying areas (<200 metres above sea level), intermediate (200-600 m) and upland (or mountain) site-types $\geq$ 600m).

Table 5.7 shows considerable differences in the performance of the model in these three zones. In the high altitude zone, it explains 80% of O_3 variability, though with a relatively high RMSE (9.0 $\mu g/m^3$). At sites in the intermediate and lowland category, it explains ~50% of the variability, though the RMSE is lower. One reason for this might be that the O_3concentrations are more variable in upland areas, where photochemical activity is generally high, but where marked variations in altitude may occur over short distances. In lowland areas, in contrast, O_3 concentrations vary much less: the variation is therefore more subtle and difficult to model (resulting in a lower R^2) but the magnitude of the errors is relatively small (lower RMSE).

Table 5.7 Performance of the global LUR within different altitude ranges

Altitude range	No. of sites	Altitude mean	Adj. R^2	RMSE
<200m	581	70	0.50	7.01
200-600m	260	353	0.48	7.80
>= 600m	138	891	0.80	9.01

It should also be noted that these results are similar to those reported in Tucson (Deim and Comrie, 2002). Sites there were generally >=600 metres above sea level and the regression model gave R^2=0.9 and RMSE=9$\mu g/m^3$.

5.4.2.5 Variation by land use

Marked differences may be expected in O_3 concentrations, and thus in model performance, between urban and rural sites. Results were therefore compared between urban and rural areas, categorised according to the site descriptions in the Airbase database. A random sample of 150 sites was selected from each category (i.e. urban and rural). Table 5.8 shows the performance of the model in these two categories.

Table 5.8 Performance of the global LUR by urban vs. rural

Site category	No. of sites	R Square	RMSE
Rural	**150**	**0.71**	**7.80**
Urban	**150**	**0.58**	**6.10**

As can be seen, the model explains a higher percentage of variability in O_3 concentrations in the rural sites (R^2 = 0.71) compared to the urban (R^2=0.58), though the RMSE shows a reverse pattern. These results are comparable to those from universal co-kriging reported by Beelen at al. (2009), which gave R^2 = 0.64 for rural sites and 0.59 for urban, with an RMSE of 7.75 and 5.59 $\mu g/m^3$ respectively. The general improvement in the LUR model in this study is probably due to using higher resolution input data and a larger number of monitoring sites.

5.4.2.6 Conclusions of post hoc studies

All these post hoc analyses suggest that site-type or area-specific models might well work better in some circumstances, though the gain in accuracy on the basis of the results obtained here may not be large. Nevertheless, the limitations of such models need to be recognised. The number of sites is not large, and some site types and areas are only poorly represented in the O_3 data, so models developed in this case may be poorly calibrated. The use of area or site-type specific models is also likely to lead to marked discontinuities at the boundaries of the areas. On this basis, the final global LUR model developed here can be considered to offer a sound foundation on which to build space-time models of O_3 concentrations, for use in exposure assessment.

5.5 Summary

In summary, the underlying principle of spatial modelling is to attempt to model the three elements of variability: global trend, random spatially correlated variation, and noise. Most of the spatial variation in O_3 concentrations is related to the random spatially correlated variation element.

Altitude, local road length, summer temperature, and high density residential land within a 1 km window, windspeed, topex and distance to sea were found to be the most significant predictors in the spatial model. These variables were used as proxies for the distribution of emission precursors (NOx and VOC). Meteorological data on solar radiation and temperature were used to represent the capacity for photochemical activity. Topographic data, on altitude and topographic exposure (topex), along with wind speed, were used to represent the potential for local or regional-scale transport, deposition, and chemical reactions all of which vary in response to meteorological conditions and the terrain. For dispersion processes, both total precipitation and agriculture were used as indicators of wet and dry deposition, respectively. Some variables used in the modelling also act as proxies for a number of different factors and processes: distance to sea, for example, provides a proxy for marine-derived O_3, for transport of O_3 either on- or off-shore by sea breezes, and for the regional-scale effect of the sea on meteorological conditions and photochemical activity.

The success of the spatial model, evidenced by the external validation, and comparison with the few previous attempts modelling the long term concentrations at the continental scale , emphasises that LUR is an appropriate technique to a derive high resolution (100m) map of long term O_3 concentrations across Western Europe. The spatial model explained 67% of O_3 variation over six years, from 2001 to 2007, with an RMSE = 7.6 $\mu g/m^3$. This spatial model is the first to estimate the long term O_3 concentrations across Europe at such a fine spatial resolution (i.e.100m), and as such at a scale suitable for semi-individual exposure assessment in epidemiological studies and/or HIA. It will further be combined with the temporal models produced using Fourier analysis (in chapter 6) to produce space–time models for sub-areas within the spatial domain of Western Europe.

6 Temporal models

As mentioned in Section 2.3, the systematic temporal variability in O_3 concentrations was modelled using Fourier analysis to develop time functions for each site type. This chapter explains how the time functions were generated and then combined using regression analysis to produce the temporal model for each site type. Each temporal model was subsequently applied to unmonitored target locations by weighting them by the probability of occurrence of each site type at that location. Probabilities were produced using MLOR. The overall steps in this procedure are summarised in Figure 6.1.

6.1 Introduction

As outlined in Chapter 5, spatial variation in phenomena such as pollutant concentrations can be characterised as comprising three main components of variation: spatial trend or drift, random spatially correlated variation, and random spatially uncorrelated variation or noise. These components provide a framework within which geostatistical techniques for spatial modelling (i.e. kriging) have been developed and are also an encompassing framework for LUR, as used in this research.

The same categorisation can usefully be extended to temporal variations. These, likewise, can be seen to comprise three components. The first is systematic variability: i.e. patterns that are systematic and repeated over time. The second can be referred to as random, temporally correlated variability. This comprises variations associated with temporally varying factors which behave non-systematically. Finally there is the totally random variation, or noise. In the temporal, as in the spatial variation, each of these components of variation may operate at different scales. In the temporal case, these range from the very short-term (e.g. over periods of seconds or minutes) to the very long-term (e.g. over decades or centuries). The scales of most relevant here are the intermediate scales – i.e. diurnal patterns (from one hour to another within a day, or from one part of the day to another); hebdomonal (across a week); and seasonal.

In the case of O_3, the systematic variation largely reflects the repeated natural variation of chemical reactions associated with seasonal or diurnal variations in temperature, solar radiation, wind speed and wind direction. Random temporally correlated variability relates to less consistent temporal

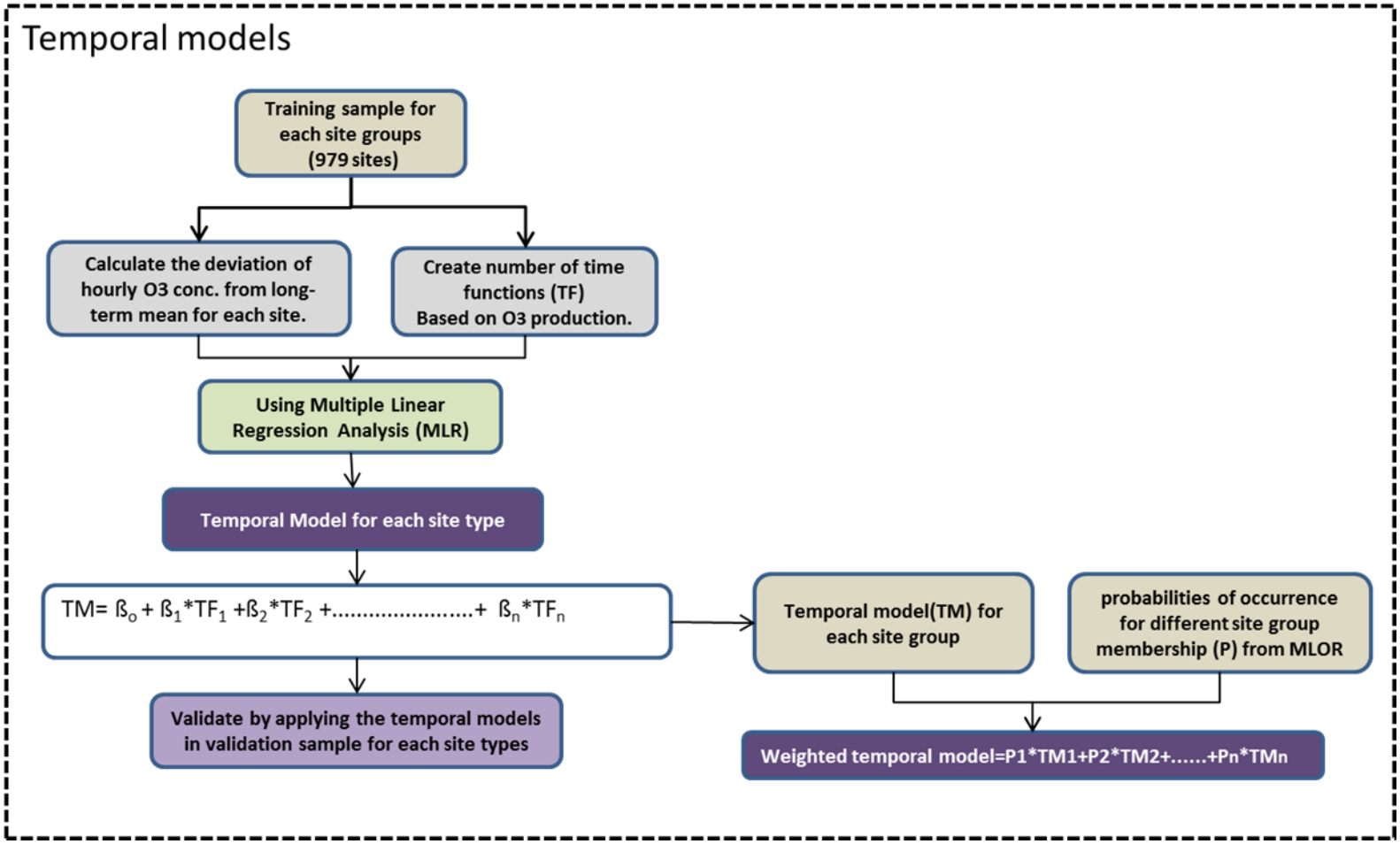


Figure 6.1 Outline of the procedure to create time function models for site types

variations in the factors controlling O_3 formation, transport and destruction, such as short-term weather events and emission activities. The noise component is due to unpredictable, seemingly random effects, such as fluctuations in traffic flows or local wind gusts, or more regional scale storms and heat episodes, as well as measurement error. Some of the apparent noise may also be due to inputs of O_3 from other areas, as a result of long-range transport.

In reality, the three overlap and are not wholly distinct. Some of the systematic variation described above, for example, is due to cycles of human activity which, although not wholly predictable, show some degree of regularity when averaged. It is useful, however, to model these components of variability separately, for they pose different challenges and require different data and methods. In this chapter, the focus is on the systematic variability, which is modelled using Fourier analysis. The random temporally correlated variability is covered in Chapter 7 by adding meteorological factors into the analysis. Noise, obviously, cannot be explicitly modelled – though the overall degree of noise that exists in any system can be roughly estimated – and will be indicated by the error in the modelling.

6.2 Components of temporal variability in O_3 concentrations

As outlined above, the approach to modelling adopted here is to describe the systematic variations in O_3 concentrations, using Fourier analysis. This approach is based on the assumption that O_3 exhibits systematic periodic behaviour, especially over the day and across seasons (Duenas et al., 2004, Coyle et al., 2002, Nolle et al., 2002, Böhm et al., 1991). This reflects the systematic variation in the different factors that control O_3 concentration (e.g. photochemical reaction, meteorological factors, quantity of precursors).

6.2.1 Seasonal variation

As has been implied, the seasonal variability is related largely to meteorological factors (temperature, wind speed, sun duration, and total precipitation). They thus provide the background onto which shorter term variations are imprinted. Other studies claim that the annual variation in stratosphere-troposphere exchange also contributes to this seasonal pattern (Levy et al., 1985, Logan, 1985). Monks (2000), however, claims that there is no seasonal variability in stratosphere-troposphere exchange.

The summer season is characterised by high temperatures and longer duration of sunshine hours, which lead to active photochemical production of O_3, and high concentrations. In contrast, in winter as temperature and sunshine decline, and precipitation increases (thereby facilitating wet deposition), O_3 production falls and concentrations are lower. The variation between summer and winter increases as the temperature difference rises; it thus implies more marked variation in more continental climates, or in Polar Regions where there are strong differences in solar radiation between summer and winter.

In reality, the picture is more complex than this. The seasonal pattern of O_3 variability – characterised by a broad spring or spring-summer O_3 maximum in the northern hemisphere (Fernández-Fernández et al., 2011, Vingarzan and Taylor, 2003, Nolle et al., 2002, Monks, 2000, Mayer, 1999) – depends on spatial location. The higher levels of summer solar radiation and temperature, and lower wind speeds, seen in southern countries, create circulation patterns linked to the diurnal flux of sea breezes, and generate a reservoir for O_3 and its precursors. This leads to high photosmog episodes in cities and coastal regions (Nolle et al., 2002). Across Europe, therefore, the amplitude of seasonal variation tends to increase in a north-west to south-east direction, and also shifts the maximum to late summer in the southern countries (Scheel et al., 1997).

This seasonal pattern of variability is seen most clearly in remote inland rural sites, unaffected by local emission sources. In these, typically, O_3 concentrations show a marked seasonal contrast, with a clear spring maximum, and little or no daily variability (Tarasova and Karpetchko, 2003, Monks, 2000). In coastal areas, however, variation is less because the temperature range tends to be reduced due to winter warming of the air by the sea, and cooling in summer. Thus the period of maximum O_3 concentrations tends to be broader, and extend from spring into summer with the spring maximum higher than the summer one (Fernández-Fernández et al., 2011). A similar pattern of a broad spring-summer maximum is likely to be shown in urban sites (Fernández-Fernández et al., 2011, Vingarzan and Taylor, 2003, Wang et al. 1988a cited in Monks, 2000), although in these cases it has been observed that the summer maximum tends to be slightly higher than that in spring (Mayer, 1999) (Figure 6.2). Wang et al. (1998a) explanied that this pattern is due to two factors: firstly that the life time of O_3 is longer in winter/spring; secondly active high-level transport of O_3 from remote sites. Simpson (1995), however, related this pattern to enhanced photochemical reaction by the accumulated O_3 precusors built up during winter.

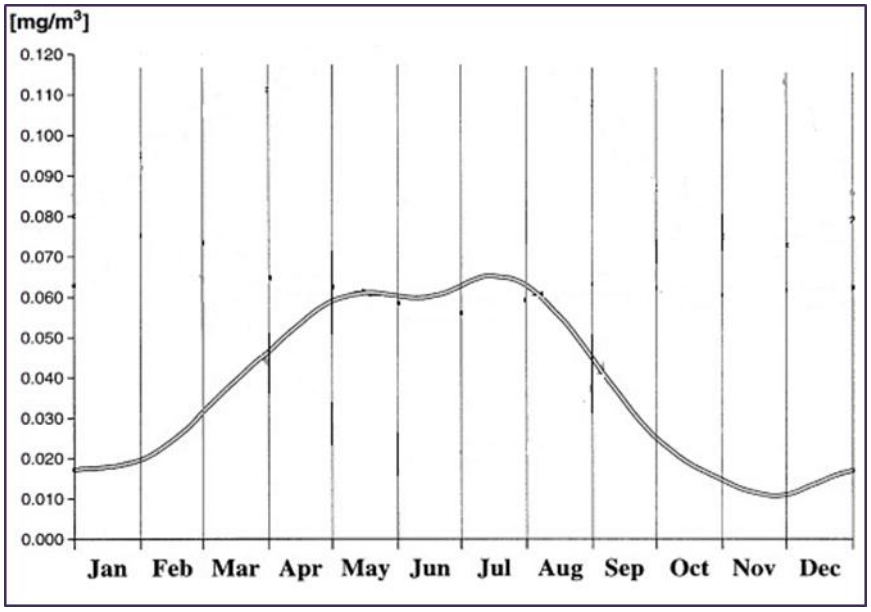


Figure 6.2Average seasonal variation of O_3 at Stuttgart-Bad Cannstatt (an urban site) for the period 1981-1993 (from Mayer 1999)

6.2.2 Hebdomadal and diurnal variation

Adding local sources to the pattern generates short term variations, represented by day-of-week (hebdomadal) and hour of day (diurnal) effects, both associated with variations in emission intensity. Diurnal effects also result from differences in O_3 production and decay between day and night. In each case, the patterns vary spatially - i.e. between the major types of environment – north vs. south, coastal vs. inland and urbanised vs. rural.

O_3 shows a weekly cycle from one day to another, that broadly reflects the pattern of human activities over the week. Concentrations tend to be higher at the weekend, and especially on Sunday compared to weekdays (Jenkin et al., 2002, Marr and Harley, 2002, Pont and Fontan, 2001, Wilby and Tomlinson, 2000). Smaller variations may also be seen between weekdays. These cycles are produced by variations in the local sources of O_3 precursors, especially NOx emitted from traffic and industrial activities. For instance, volumes of traffic often rise on Mondays, when people go back to work, fall off slightly from then until Thursday and peak again on Friday, then reach a minimum over the weekend. This leads to a stronger weekend contrast in urban sites compared to rural sites, and little or no hebdomadal effect in remote sites (Jenkin et al., 2002, Mayer, 1999).

A strong diurnal cycle of O_3 concentrations can be expected between day and night due to differences in the rates of photochemical activity, as outlined earlier Section 2.1.1. During the daytime, solar radiation is high, which activates photochemical reactions to form O_3 at a more rapid rate than its destruction. In contrast, O_3 concentrations are reduced at night due to the lack of photochemical activity. This pattern will vary, however, depending on the spatial context (e.g. urban, rural, industrial, remote area or traffic).

Superimposed on this pattern is the effect of the rate of supply and destruction of precursor pollutants. The ratio of NO:NO_2 is especially important in this respect, for while NO acts to destroy O_3 (by conversion to NO_2), NO_2 tends to promote O_3 production, by dissociating to NO and O.

NO and NO_2 show strong diurnal patterns, largely due to variations in emission from road traffic and some industrial sources. NO typically has two peaks within the day: one in the morning and another in the evening, associated with periods of heavy traffic (Sanchez and Sanz, 1994). As a secondary pollutant, NO_2 concentrations lag behind NO, and often peak later in the morning and evening. O_3 concentrations can be expected to lag somewhat further behind, peaking when the NO:NO_2 ratio is lowest and reaching a nadir when this ratio is highest, as shown in Figure 6.3. These patterns vary, however, both with distance from NO source, and, as has been indicated already, with levels of solar radiation.

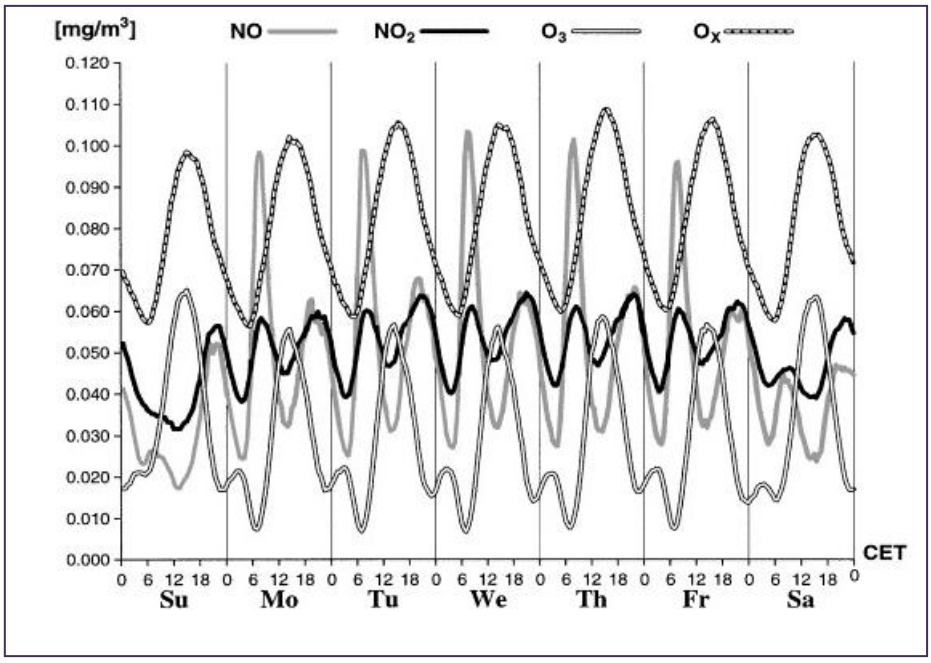


Figure 6.3 Average weekly and diurnal cycle of NO, NO_2, O_3, and O_x at Stuttgart-Bad Cannstatt (an urban site) for the period 1981-1993 (from Mayer 1999)

Because of these factors, urban area tend to show high diurnal variability in O_3 concentrations compared to more remote areas (i.e. rural areas) (Böhm et al., 1991). In principle, an urban area can be expected to exhibit a curve which peaks in the afternoon and decreases at night (Finlayson-Pitts and Pitts Jr, 1986). In large and densely populated urban areas with high traffic volumes, the pattern will be accentuated, leading to marked differences between morning, afternoon, and night-time concentrations. The highest peak occurs when solar radiation is high, during the afternoon times, and then declines as the effect of reduced solar radiation and increased traffic emission take effect during the early evening (Figure 6.3). The same occurs, albeit to a lesser extent, during the morning cycle. The more intense the road traffic, the deeper the trough in O_3 concentrations. If the urban site is located downwind, the cycle is shifted later in the day, due to the lag caused by dispersion (Finlayson-Pitts and Pitts Jr, 1986). A low maximum concentration in early morning (between ca.12.00 – 4.00 am) results from downwind transport of O_3 from places with high concentrations; in rural this variation is absent (Mayer, 1999), as illustrated in Figure 6.4.

In remote, rural areas, there are likely to be very few scavengers of any sort, and little input of O_3 by dispersion from elsewhere, so the pattern is driven simply by photochemical reactions. Variations also occur with altitude, so the amplitude of variation is height-dependent: at high altitudes, variations in O_3 concentrations are driven mainly by air mass recirculation (Millán et al., 2000). In general, coastal and upland sites have similar diurnal patterns of O_3 due to the absence of any marked reduction at night, as occurs in lowland and urban sites (Sundberg et al., 2006), as shown in Figure 6.5.

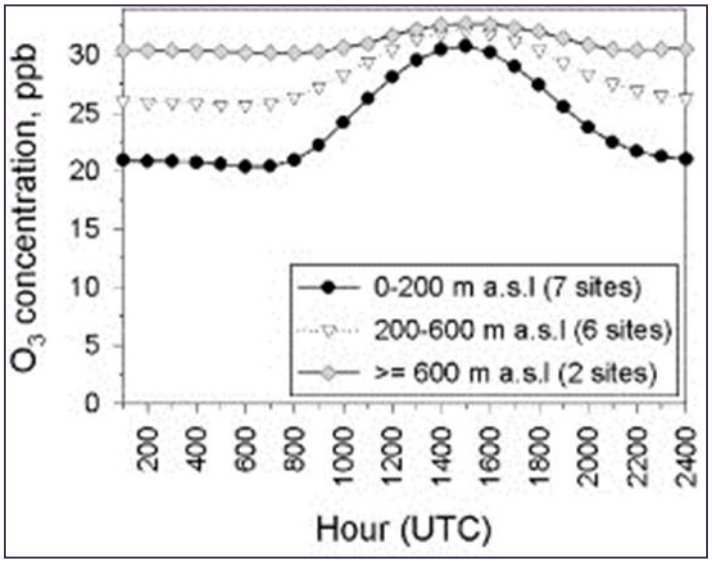


Figure 6.4 Typical diurnal cycles at rural sites, averaged into groups by site altitude (from Coyle et al., 2002)

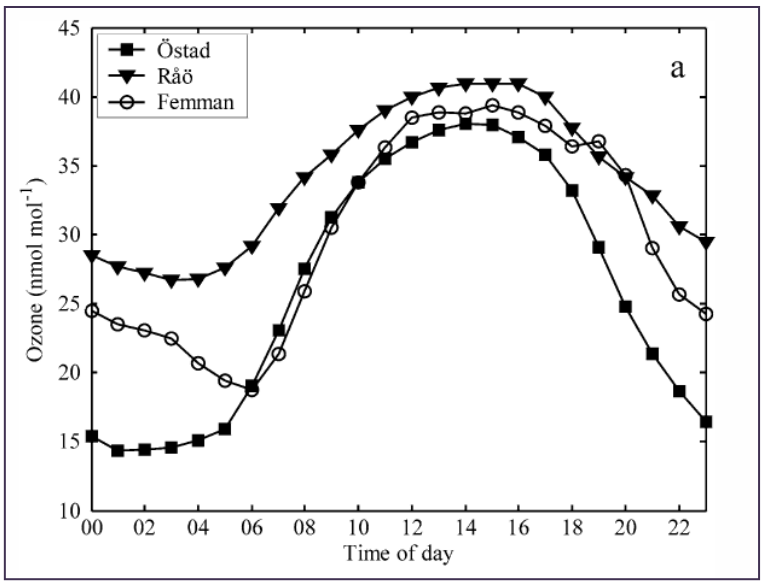


Figure 6.5 Typical diurnal variation in O_3 concentration at three sites: Ostad (rural site located in a broad valley), Rao (coastal site), and Femman (urban site) in Sweden during 2004 (from Sundberg et al., 2006)

6.2.3 Implications for modelling

From all these considerations, it can be argued that O_3 concentrations are likely to show systematic cycles over three main time periods: seasonally, weekly (hebdomonally) and diurnally. Each of these varies in its intensity (i.e. the amplitude of the peaks and troughs), duration, timing and frequency depending on their geographic (and especially meteorological) context and the associated patterns of emission intensity. This systematic temporal behaviour in O_3 concentrations can, in principle, be described by a set of functions of time. Insofar as the patterns are consistent across any specific type of environment (i.e. site type) these functions then provide the means of predicting the temporal variability at unmonitored locations.

The functions themselves can be defined in a number of ways. One is by decomposing the O_3 concentrations into a set of functions using ARIMA or similar approaches. This, however, is likely to result in complex, possibly over-fitted functions, which are not necessary interpretable or transferable to other sites, as discussed in Subsection 2.2.3.2. Fourier analysis provides an alternative, and has the advantage of providing semi-deterministic control to the analyst (who can define the functions on the basis of a priori knowledge). It is widely used in technological fields that require methods to describe and model time-varying phenomena such as radio-communications. It has also, however, begun to be used in several areas of environmental science, including studies of water quality and atmospheric chemistry (Skene et al., 2010, Richards and Baker, 2002, Damsleth and Spjotvoll, 1982). Here, there is scope to use the approach to define the different systematic patterns that occur in O_3 signatures as a result of determinants such as variations in solar radiation, temperature and associated photochemical processes, as well as regular variations in emission intensity.

6.3 Principles of Fourier analysis

Fourier analysis, as summarised by Piegorsch and Bailer (2005), involves the use of trigonometric functions, typically based on sine waves, to model the periodicity of time series data.

The sine wave angle is measured in radians (θ), starting from $\theta = 0$ where the angle in degrees equals 0°, and ranging up to 2π where the angle equals 360°.

This time function can be built in form of**:**

$$f = A \sin(2\pi(t_i - \theta)/p) \qquad \text{Equation 6-1}$$

where p is period of interest (e.g. 24 hours, 7 days); t represents the target time within that period; A is the amplitude of the wave, which is defined by the maximum height of peak or the depth of the trough relative to the basal axis; and θ is the phase angle that lags the wave either to the right or to the left.

Two simple functions, using 24 hours of day as the period, are shown in Figure 6.6. These define a sine wave, $f = \text{Sin}[2\pi t/24]$, and a cosine wave, $f = \text{Cos}[2\pi t/24]$. Variations of this simple function can be created by shifting the wave horizontally, by setting θ to a positive or negative value depending on the desired direction of lag (H). For example, setting θ to +2 such that $f = \text{Sin}[2\pi[t+2]/24]$ shifts the sine wave in Figure 6.6 two units to the left (Figure 6.7). Equally the wave can be shifted vertically by the inclusion of a constant: $f = \text{Sin}[2\pi t/24] + 0.5$ shifts the wave 0.5 units upwards, as illustrated in Figure 6.7.

By using these features to manipulate the waves, it is possible to recreate practically any pattern to match the temporal behaviour in the target variable, for different time periods of interest. The goal here is to produce the best approximation of O_3 concentrations for different time periods (seasonal, weekly, diurnal) and for different spatial locations (urban, rural, coastal, remote, etc.), on the basis of the principles outlined above.

By using one time function, a simple model can be built, as shown in Equation 6.2. This, however, usually needs to be calibrated to the empirical data using regression analysis (Equation 6.2). This approach assumes that there is no temporal correlation within the error term ε and that the period is known. Regression analysis essentially weights the function, and thus increases or reduces the amplitude of the effect:

$$Y_i = a + \beta_1 f_1 + \varepsilon_t \qquad \text{Equation 6-2}$$

where Y_i is the observed time series dataset, a is the height of the wave, β_1 the coefficient representing the amplitudes of f_1, and ε is an error term.

In many cases, time series data cannot adequately be described by a single, simple function. In this case, different functions need to be combined, additively, to represent the different periodicities in the data, or asymmetry in the waves. This is done through a multiple Fourier regression analysis, as shown in Equation 6.3.

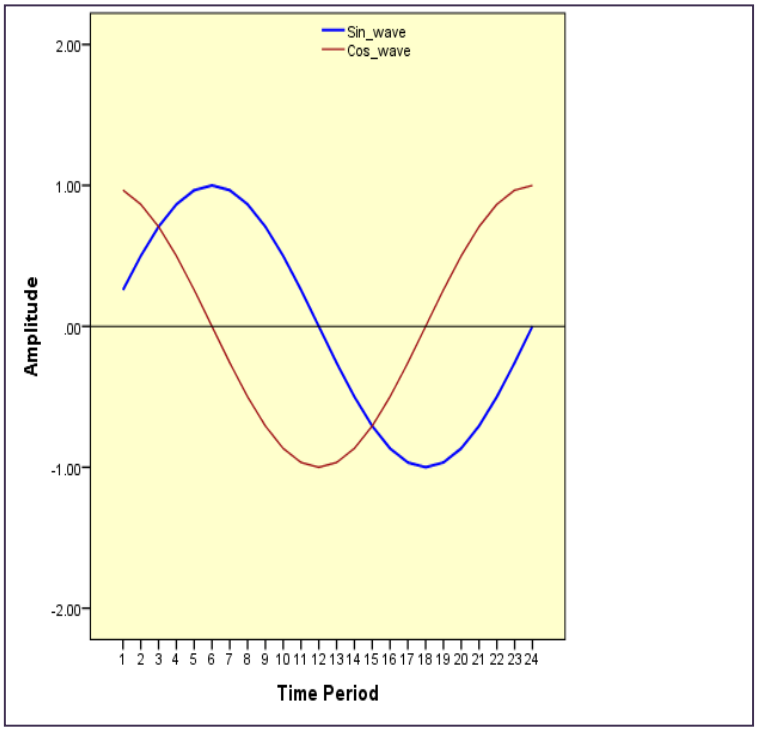


Figure 6.6 Plots of simple time functions: sine and cosine

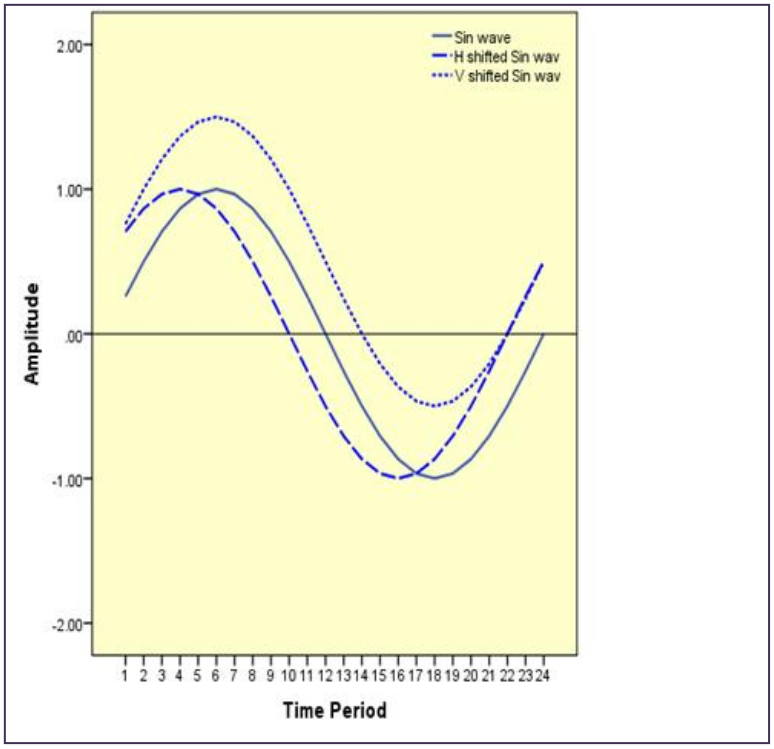


Figure 6.7 Shifted time function

$$Y_i = a + \beta_1 f_1 + \beta_2 f_2 ... + \beta_n f_n + \varepsilon_t \qquad \textbf{Equation 6-3}$$

where Y_i is the observed time series dataset, a is the height of the wave, $\beta_1$and β_n the coefficients representing the amplitudes of the n waves, and ε is an error term.

As the Fourier analysis model can only represent the systematic variation, other covariates representing the unsystematic (non-periodic) variation also need to be considered in order to reduce the white noise (i.e. volume of error). As the variation in O_3 concentrations is affected by meteorology (see Section 3.2.3.5), it is likely that meteorological variables provide the best basis for capturing the non-systematic but temporally correlated variability in a temporal model. This is addressed in Chapter 7.

6.4 Methodology

Temporal models for each site type were built by first generating time functions to represent the pattern of O_3 behaviour over different time periods. Linear regression analysis was then used to combine the functions into a time model for each site type, using standardised hourly concentrations, created by subtracting the site long-term mean concentration from each hourly concentration of the site (i.e. deviation from the long-term mean O_3 concentration). Thirteen temporal models were thus developed, one for each of the site-types.

6.4.1 Development of time functions

Different approaches can be used to develop trigonometric functions to describe a time series of data. One approach is to assume that the underlying systematic patterns run uniformly throughout the data set, and thus to identify the patterns by examining averaged data (Barnett, 2004). In this case data are averaged across the whole study period for each basal averaging period (e.g. hours of the day in order to identify a function describing diurnal patterns). The averages can then be plotted (e.g. for each of the twenty four hours in a day) and examined to identify the shape and timing of the underlying patterns.

An alternative is to proceed deterministically, and create functions which describe the temporal signal for a variety of possible scenarios (i.e. signatures), based on theory. In this case, this would involve specifying the general patterns of O_3 concentrations expected both over different time

periods (a day, a week, a year) and for different situations (e.g. more or less urbanised, upland and lowland, northern and southern).

The latter approach was adopted here, as it is more generalized and models are not built for specific sites. A series of functions was created to represent fluctuations over daily, weekly and seasonal time scales, based on the principles outlined in Section 6.1. Time functions were computed to match each of the site types identified in Chapter 4.

Functions were generated by applying the following steps.

1. Create a number of simple independent functions for seasonal and daily function peaked at different times and for an appropriate basal averaging period.
2. Create different versions of these by shifting them by 1 increment forward or backward in time (i.e. for different time lags).
3. Enter the resulting functions into a multiple regression analysis, using hourly standardised (the deviation from the mean) O_3 concentration for all sites within the site type, as the dependent variable.

As an illustration, to model the diurnal variations in O_3, the procedure was applied as follows:

A. Using the hourly data (i.e. with an hour as the basal unit of time), a simple 24-hour day function, a 'starter function', was defined giving a peak at 13:00 hours, and termed D13 (D defining it as a day function and the number representing the peak hour).
B. This was then shifted by one-hour increments to create a family of lagged functions with peaks at different times of the day to represent day and night-time cycles. Figure 6.8, shows functions for the first five diurnal functions.
C. A second family of 'complex' (i.e. double-peaked) function was created to represent more complex patterns during the day. In this case, the starter function was defined with two peaks at 12:00 and 24:00, and then these lagged by one hour to create different versions, Figure 6.9.

In the case of the seasonal pattern, the same approach was applied using daily averages (i.e. with the day as the basic time unit). The starting function was designed to peak at the beginning of May (spring season) and this then shifted by 10 days at a time, to create a family of seasonal functions, Figure 6.10.

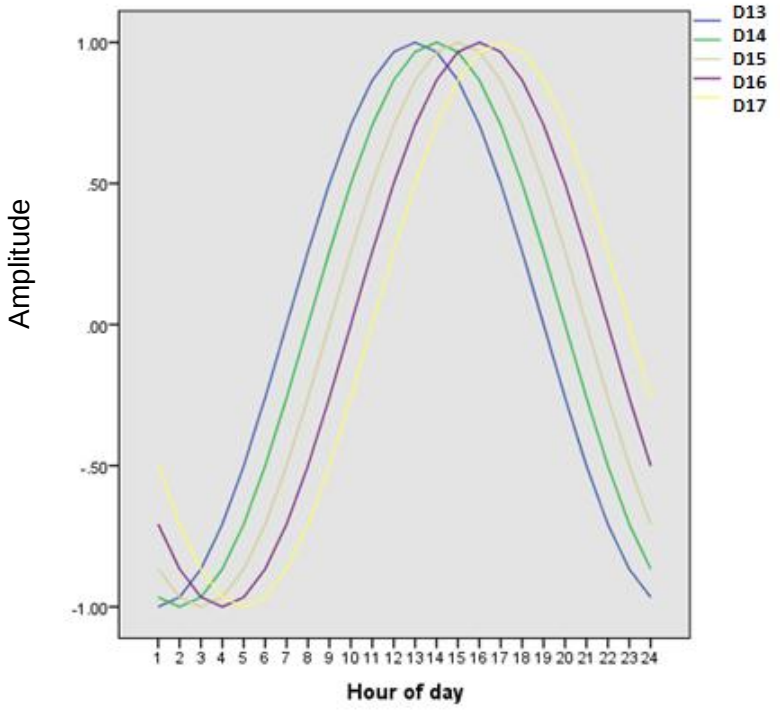


Figure 6.8 Simple diurnal time functions showing an afternoon peaked from 13.00 to 17.00

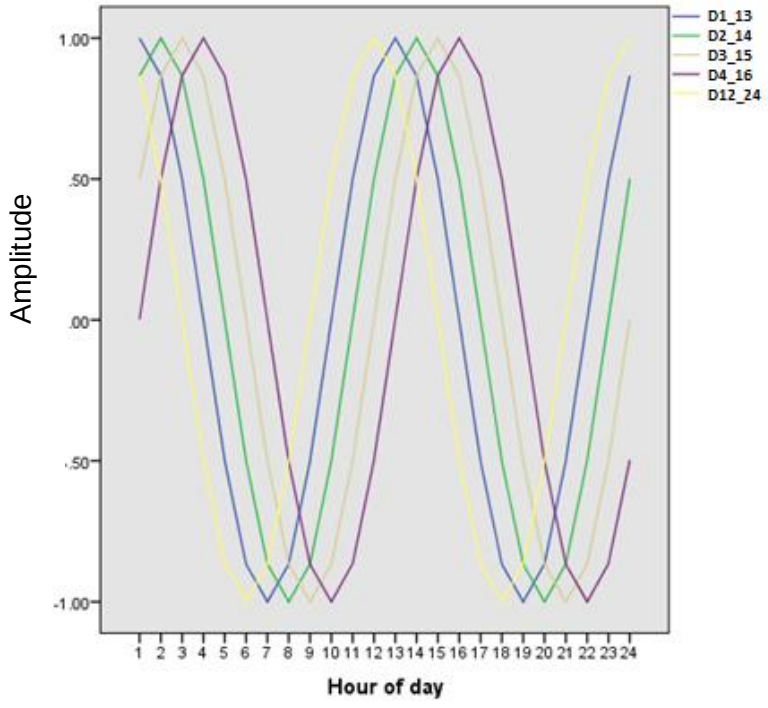


Figure 6.9 Complex diurnal time functions with a double-peak in the early morning and afternoon

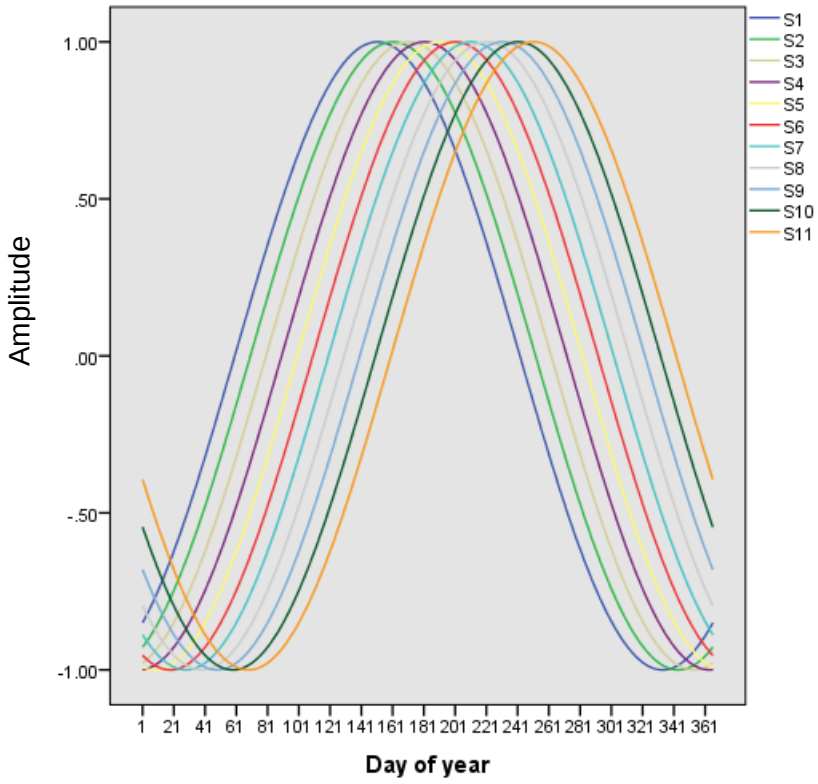


Figure 6.10 Seasonal Time functions for S1 and its sequence versions

In the case of the day of the week, it was felt that a simple function could not easily be constructed, because of the disjunct nature of the effect – i.e. weekend days are fundamentally different in terms of emissions from working days. Therefore, day of week was coded as three binary variables (Sunday, Saturday, and weekday) and these entered as dummy variables into the regression analysis. In this way the regression analysis determined the size of the additive effect for each of these day-of-week periods. The 32 functions thus created are summarised in Table 6.1.

Table 6.1 Description of seasonal and daily generic time functions

Time period	Time Function	Description
Day of Year [A]	S1=Sin(2*3.142*(DoY-60)/365)	Seasonal function with one wave peak timed in May
	S2=Sin(2*3.142*(DoY-70)/365)	Seasonal function with peak shifted by 10 days from S1
	S3=Sin(2*3.142*(DoY-80)/365)	Seasonal function with peak shifted by 20 days more from S1
	S4=Sin(2*3.142*(DoY-90)/365)	Seasonal function with peak shifted by 30 days from S1
	S5=Sin(2*3.142*(DoY-100)/365)	Seasonal function with peak shifted by 40 days more from S1
	S6=Sin(2*3.142*(DoY-110)/365)	Seasonal function with peak shifted by 50 days from S1
	S7=Sin(2*3.142*(DoY-120)/365)	Seasonal function with peak shifted by 60 days from S1
	S8=Sin(2*3.142*(DoY-130)/365)	Seasonal function with peak shifted by 70 days more from S1
	S9=Sin(2*3.142*(DoY-140)/365)	Seasonal function with peak shifted by 80 days from S1
	S10=Sin(2*3.142*(DoY-150)/365)	Seasonal function with peak shifted by 90 days from S1
	S11=Sin(2*3.142*(DoY-160)/365)	Seasonal function with peak shifted by 100 days from S1
Hour of Day [B] with one Peak	D13=SIN(2*3.142*(HoD-7)/24)	Daily function peak: at 13.00
	D14=SIN(2*3.142*(HoD-8)/24)	Daily function peak: at 14.00
	D15=SIN(2*3.142*(HoD-9)/24)	Daily function peak: at 15.00
	D16=SIN(2*3.142*(HoD-10)/24)	Daily function peak: at 16.00
	D17=SIN(2*3.142*(HoD-11)/24)	Daily function peak: at 17.00
	D22=COS(2*3.142*(HoD+2)/24)	Daily function peak: at 22.00
	D23=COS(2*3.142*(HoD+1)/24)	Daily function peak: at 23.00
	D24=COS(2*3.142*(HoD)/24)	Daily function peak: at 24.00
	D1=COS(2*3.142*(HoD+23)/24)	Daily function peak: at 1.00
	D2=COS(2*3.142*(HoD+22)/24)	Daily function peak: at 2.00
	D3=COS(2*3.142*(HoD+21)/24)	Daily function peak: at 3.00
	D6 =COS(2*3.142*(HoD+6)/24)	Daily function: trough at 6.00
	D7 =COS(2*3.142*(HoD+5)/24)	Daily function: trough at 7.00
	D8 =COS(2*3.142*(HoD+4)/24)	Daily function: trough at 8.00
	D9 =COS(2*3.142*(HoD+3)/24)	Daily function: trough at 9.00
	D10 =COS(2*3.142*(HoD+2)/24)	Daily function: trough at10
Hour of Day with two Peaks and trough	D12_24=SIN(4*3.142*(HoD+3)/24)	Daily function peak: 12 and 24.00
	D1_13=SIN(4*3.142*(HoD+2)/24)	Daily function peak: 13 and 1.00
	D2_14=SIN(4*3.142*(HoD+1)/24)	Daily function peak: 14 and 2.00
	D3_15=SIN(4*3.142*(HoD/24)	Daily function peak: 15 and 3.00
	D4_16=SIN(4*3.142*(HoD+23)/24)	Daily function peak: 16 and 4.00

A. Day of Year (DOY) to capture the seasonal trend over a year based on daily concentrations with 365 days
B. Hourly of Day (HOD) to capture the diurnal patterns within 24 hours.
C. post hoc time functions to represent this variation

6.4.2 Regression analysis (Fourier analysis)

These generic functions have to be fitted to the data by determining the amplitude (the weight of each function) in order to create a more complex, additive function that best matches the time signature in the data. Analysis was done using regression analysis in which all the functions were entered as potential predictors and the hourly standardised (deviation from the mean) O_3 concentration used as the dependent variable. Analysis was done for each site type separately, and using only the training sites (i.e. excluding the 20% of reserved validation sites). A supervised stepwise regression analysis was run as follows:

1- Seasonal functions were entered first, followed by the three day-of-week variables (weekday, Saturday and Sunday).

2- Variables were excluded if P >0.05, VIF>5, or the increase in the adjusted $R^2 < 0.01$.

3- Diurnal functions were then offered into the model together with all included variables from the preceding step 1, and applying the same rules as in step 2.

4- Residuals from step 3 were then examined by:

   a) Plotting a boxplot for hour of day, to explore any residual diurnal pattern. Where any systematic diurnal variation was suspected, additional post hoc functions were developed in an attempt to describe it.
   b) Scatterplots for daily residual concentrations were generated to explore any remaining pattern in the seasonal systematic variation. Where this seemed possible, additional post-hoc functions were developed to describe these systematic variations.
   c) Potential post hoc functions were regressed against the residual O_3 concentration (from step 3) and were retained if conditions in step 2 were achieved.

5- Once the model was considered to be finalised, measures of the statistical goodness of fit (adjusted R^2 and RMSE) were estimated by applying the model to the validation dataset for each site type.

6.5 Results and interpretation

The first Section 6.5.1 illustrates the nature of the resulting functions using the example of two site types:

- Site type 1 (Forested hill-lands): sites are scattered over the whole study area, though mainly at large distances from the sea, and are characterised by rural land cover (especially forest) with little infrastructure; mean altitude is relatively high.
- Site type 7 (Heavily trafficked urban): sites are scattered over the whole study area and characterised by flat urban land cover with high road density.

Subsection 6.5.2 presents the results for the rest of the thirteen sites type.

6.5.1 Results: site types 1 and 7

Table 6.2 shows the models for site type 1 and 7 and the regression statistics after the first three steps in the analysis, outlined above. All included functions are significant, with P= 0.0005 and each one added 1% or more to the adjusted R^2.

Table 6.2 Regression model for site type 1 and site type 7 applying step 1-3 in the methodology

Site type	Model	Coefficient(β)	Adj.R^2	RMSE	R^2 change	p-value
Site type 1 (Forested hill-lands)	S2	23.98	.30	26.09	.30	.00
	D16	11.12	.36	24.88	.06	.00
	D2_14	3.78	.37	24.73	.01	.00
	Weekday	-1.69	.37	24.70	.00	.00
	Sunday	1.92	.37	24.70	.00	.00
Site type 7 (Heavily trafficked urban)	S2	15.43	.20	21.94	.20	.00
	Weekday	-4.05	.21	21.75	.01	.00
	Sunday	4.58	.22	21.71	.01	.00
	D2_14	7.02	.26	21.09	.04	.00
	D16	7.29	.30	20.50	.04	.00

Both site type models contain one seasonal function, peaking in the June (S2). Figures 6.11 and 6.12 show the values of S2, weighted by its relevant coefficients (23.98 and 15.43 for site types 1 and 7, respectively). The plots show that both site types display a similar pattern of seasonal variation but with different amplitudes, reflecting the degree to which the seasonal pattern is affected by meteorological conditions. The seasonal variation is notably flatter in site type 7 than in site type 1.

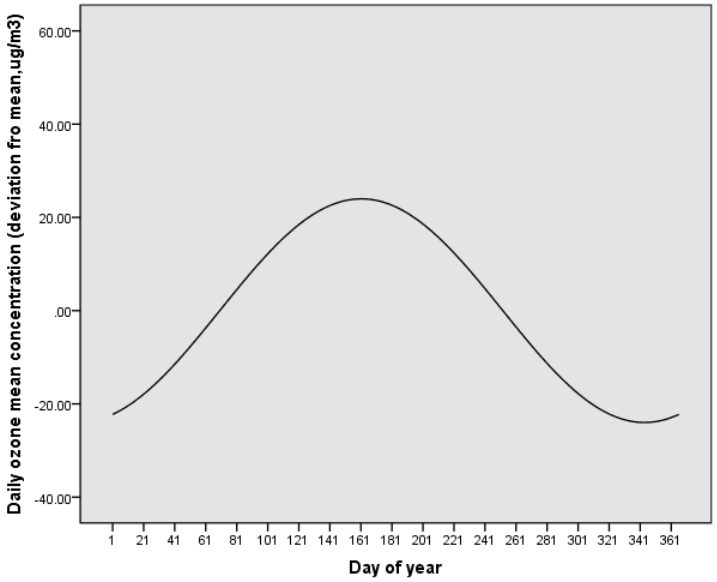


Figure 6.11 Modelled seasonal variation in site type 1 across all sites

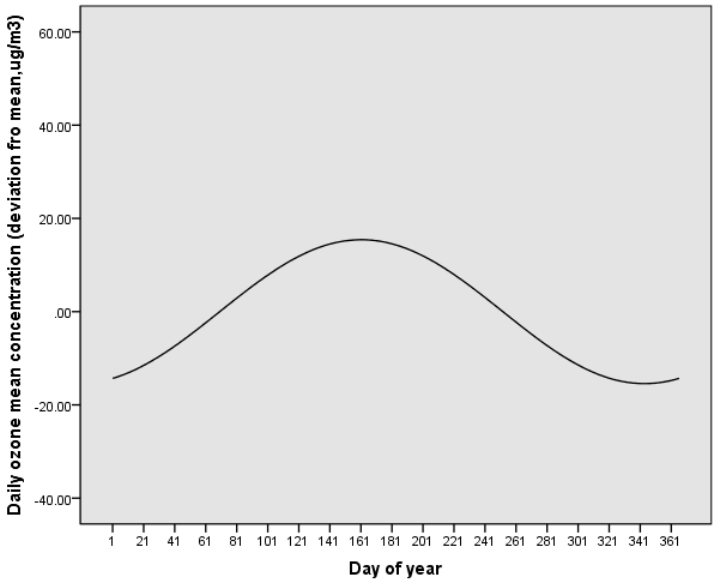
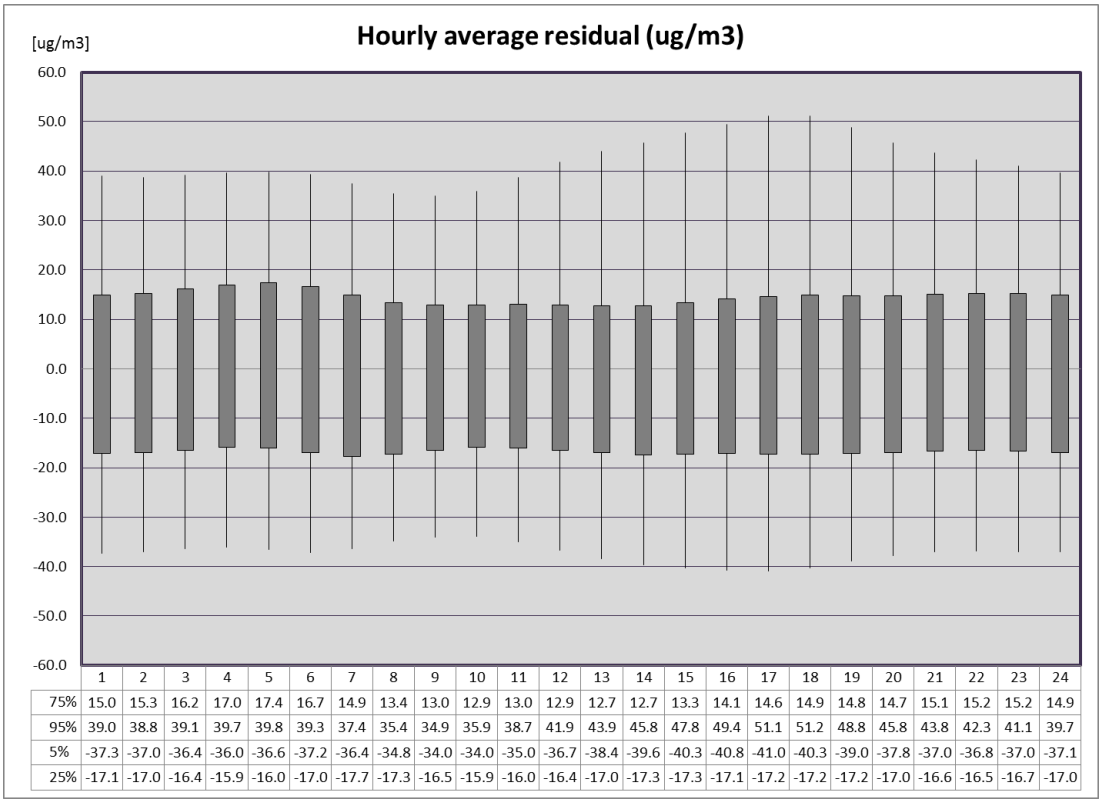


Figure 6.12 Modelled seasonal variation in sit type 7 across all sites

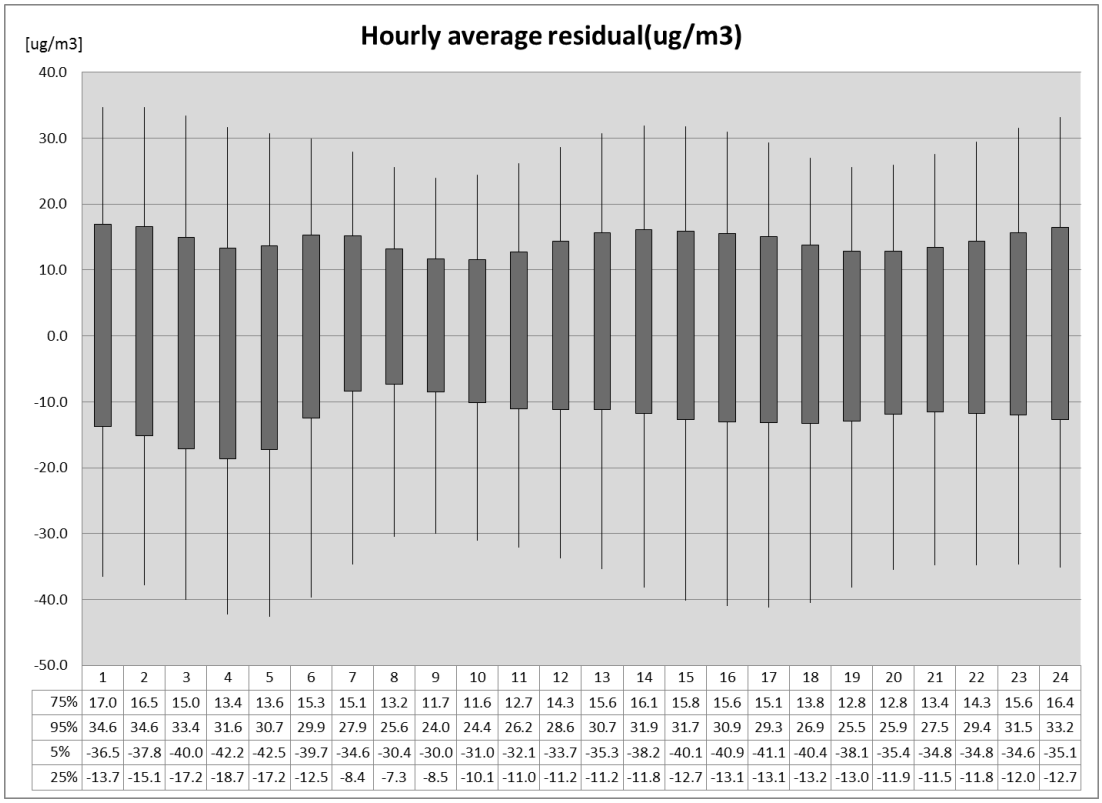
To explore whether any systematic variability remained, Figures 6.13 and 6.14 were plotted, showing boxplots for the hourly residual in the two site types. The boxplot (Figure 6.13) for site type 1 shows that the mean residual does not vary to any significant extent, but the range (variance of

the residual) exhibits some variation, reflecting differences between sites and days. Incorporating a post-hoc time function will therefore not improve the model. Site type 7 (Figure 6.14) shows some variation in the mean of the residuals, with a slight tendency for under-estimation between 02.00 and 06.00 hours. Again, however, between-hour variation is small compared to the variation within hours (differences between sites and between days), so it is unlikely that a post hoc function could improve the model to any extent.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
75%	15.0	15.3	16.2	17.0	17.4	16.7	14.9	13.4	13.0	12.9	13.0	12.9	12.7	12.7	13.3	14.1	14.6	14.9	14.8	14.7	15.1	15.2	15.2	14.9
95%	39.0	38.8	39.1	39.7	39.8	39.3	37.4	35.4	34.9	35.9	38.7	41.9	43.9	45.8	47.8	49.4	51.1	51.2	48.8	45.8	43.8	42.3	41.1	39.7
5%	-37.3	-37.0	-36.4	-36.0	-36.6	-37.2	-36.4	-34.8	-34.0	-34.0	-35.0	-36.7	-38.4	-39.6	-40.3	-40.8	-41.0	-40.3	-39.0	-37.8	-37.0	-36.8	-37.0	-37.1
25%	-17.1	-17.0	-16.4	-15.9	-16.0	-17.0	-17.7	-17.3	-16.5	-15.9	-16.0	-16.4	-17.0	-17.3	-17.3	-17.1	-17.2	-17.2	-17.2	-17.0	-16.6	-16.5	-16.7	-17.0

Figure 6.13 Boxplot for hourly residual concentration ($\mu g/m^3$) across site type 1



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
75%	17.0	16.5	15.0	13.4	13.6	15.3	15.1	13.2	11.7	11.6	12.7	14.3	15.6	16.1	15.8	15.6	15.1	13.8	12.8	12.8	13.4	14.3	15.6	16.4
95%	34.6	34.6	33.4	31.6	30.7	29.9	27.9	25.6	24.0	24.4	26.2	28.6	30.7	31.9	31.7	30.9	29.3	26.9	25.5	25.9	27.5	29.4	31.5	33.2
5%	-36.5	-37.8	-40.0	-42.2	-42.5	-39.7	-34.6	-30.4	-30.0	-31.0	-32.1	-33.7	-35.3	-38.2	-40.1	-40.9	-41.1	-40.4	-38.1	-35.4	-34.8	-34.8	-34.6	-35.1
25%	-13.7	-15.1	-17.2	-18.7	-17.2	-12.5	-8.4	-7.3	-8.5	-10.1	-11.0	-11.2	-11.2	-11.8	-12.7	-13.1	-13.1	-13.2	-13.0	-11.9	-11.5	-11.8	-12.0	-12.7

Figure 6.14 Boxplot for hourly residual concentration (μg/m^3) across site type 7

The residual from the model for site type 1 was also plotted as a scatterplot, showing the daily residual over 365 days to identify any remaining seasonal variability (Figure 6.15). In this graph, the daily residuals have been averaged across all sites and all years, to facilitate plotting: the x axis represents sequential days from January 1st to December 31st. Examination of the graph suggests that the residual (predicted minus observed concentrations) varies between different times of the year, with a tendency for the model to over-estimate during the spring months (March-May), but to under-estimate during summer (June-August). This suggests that the time functions might be amplifying the spring maximum, and implies the need for a post hoc function to describe the pattern more accurately.

A polynomial trend was therefore fitted to the graph to summarise the systematic pattern in the residual. The fit of the polynomial increased as the order was increased up to the sixth order. The function shown in Figure 6.15 is a sixth order polynomial. As with all complex polynomials, this double peak polynomial (with maxima in the spring and autumn/early winter) is not easy to interpret in terms of underlying processes. In a review by Monks (2000), however, the same pattern was reported in unpolluted locations in the northern and southern hemispheres during some

periods. The explanation appears to be that, in remote and unpolluted (low-NOx condition) regions, like site type 1, the NOx level is crucial to determination of whether the photochemical state is one of O_3 destruction or O_3 production. In low-NOx conditions, O_3 loss by photolysis (i.e. atmospheric reaction with peroxy and radicals) and surface deposition is balanced by O_3gain via entrainment from the lower free troposphere. There may also be a small additional source in summer from photolysis of nitrogen dioxide. Together, these create a cycle of winter maxima and summer minima in O_3 concentrations (Ayers et al., 1997, Ayers et al., 1992).

On this basis, the pattern observed in Figure 6.15 seems to reflect inadequacies in the seasonal time function for site type 1, so the model appears to need some adjustment. Thus an additional double-peak sine function was generated to match the polynomial: Sin [4π(DOY-100)/365]. Figure 6.16 shows the seasonal pattern that results when this is combined with the initial seasonal function for site type 1. The predicted seasonal pattern changes from one with a maximum at the end of the spring season to a broad maximum across the spring and summer, with a slightly higher concentration in summer. When this was incorporated into the model, it improved the adjusted R^2 by 1%.

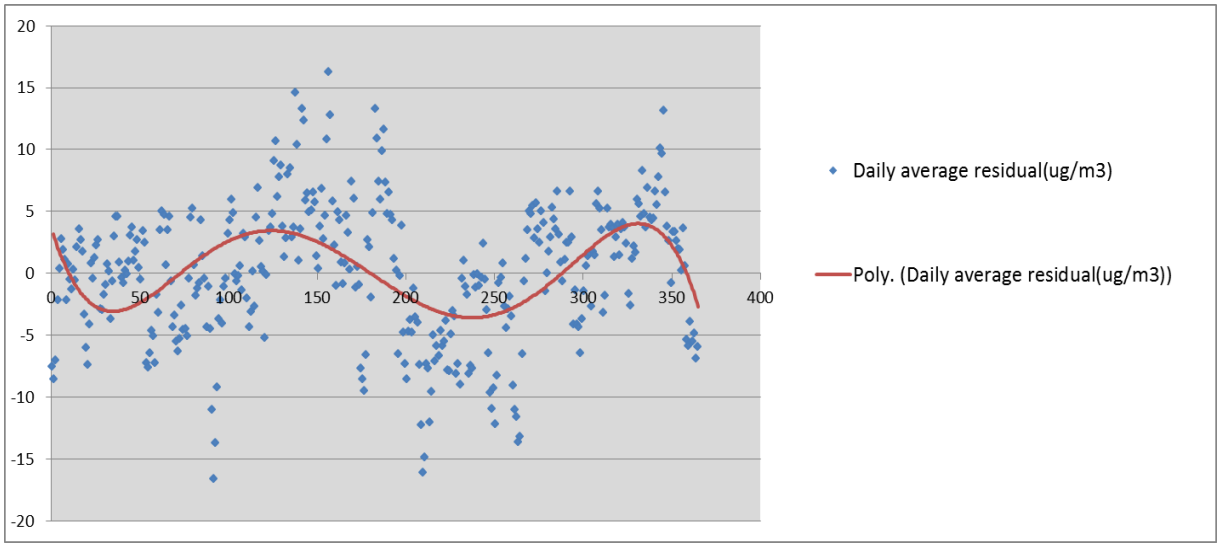


Figure 6.15 Scatterplot for daily residual over 365 days to identify seasonal variability for site type1

Residuals from the site type 7 were analysed in the same way. The polynomial function in this case suggests a small degree of underestimation in the summer months and overestimation in the winter

months (Figure 6.17). Incorporation of functions to describe this variation, however, did not improve R^2 by ≥ 1%, so the initial model was retained.

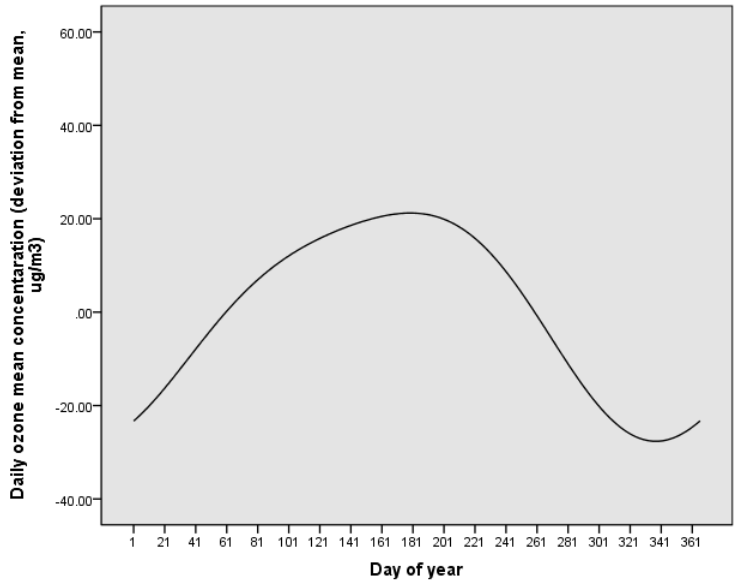


Figure 6.16 Modelled seasonal variation in site type 1 across all sites from the final model (S2+PHF1)

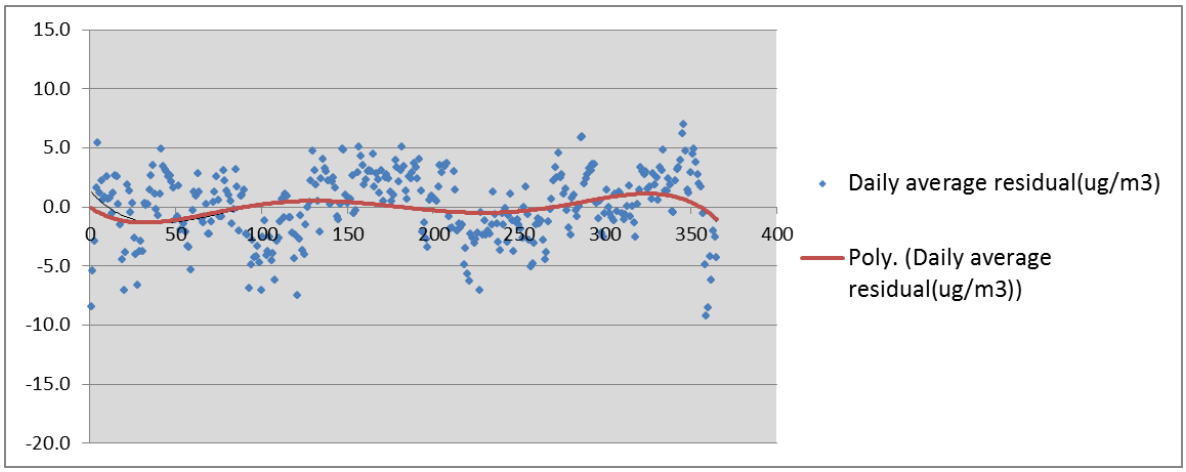


Figure 6.17 Scatterplot of the daily residual from the initial Fourier model over 365 days for site type 7

Figures 6.18 and 6.19 show the shapes of the short-term (weekly + diurnal) variation implied by the models developed for site types 1 and 7. The two site types have clearly different patterns. In site

type 1, there is a relatively simple cycle, with a peak in the afternoon and no great difference between the weekdays and weekend, as expected in remote rural areas. Site type 7 (Heavily trafficked urban) on the other hand, shows a distinct secondary peak in the night, and a marked increase in concentrations at the weekend. Both features can be thought of as typical of highly urbanised localities. The weekend effect reflects the relative lack of traffic and associated NO emissions on non-work days. The secondary peak in the night is a reflection of the following sequence of events:

1. During rush hour (from 17.00 hours onwards), NO is emitted by vehicles, which scavenges the O_3 and results in a trough in O_3 concentrations.
2. As traffic volumes subside, the NO production falls and the existing NO is transported out of the area, while NO_2 also forms, changing the NO:NO_2 ratio. O_3 production therefore gradually increases while destruction declines and a peak in concentrations occurs.
3. By the early morning, production of O_3 has fallen (due to lower photochemical activity), and traffic flows start to increase, causing a change to scavenging conditions, which reach a maximum at rush hour (ca. 08.00).
4. As traffic flows fall off, and the NO disperses, O_3 production increases rapidly, and creates a peak concentration in the mid-afternoon

It might be noted that the secondary night-time peak thus created continues into the weekend. This is largely because the same diurnal pattern has been applied in this model for every day. These data could suggest that improvements in the model might be made by applying different diurnal functions for weekdays and weekends. This was explored, by building separate diurnal models for weekday and weekend using site type 1, but the same functions were selected by the model (see Appendix B, Section VII).

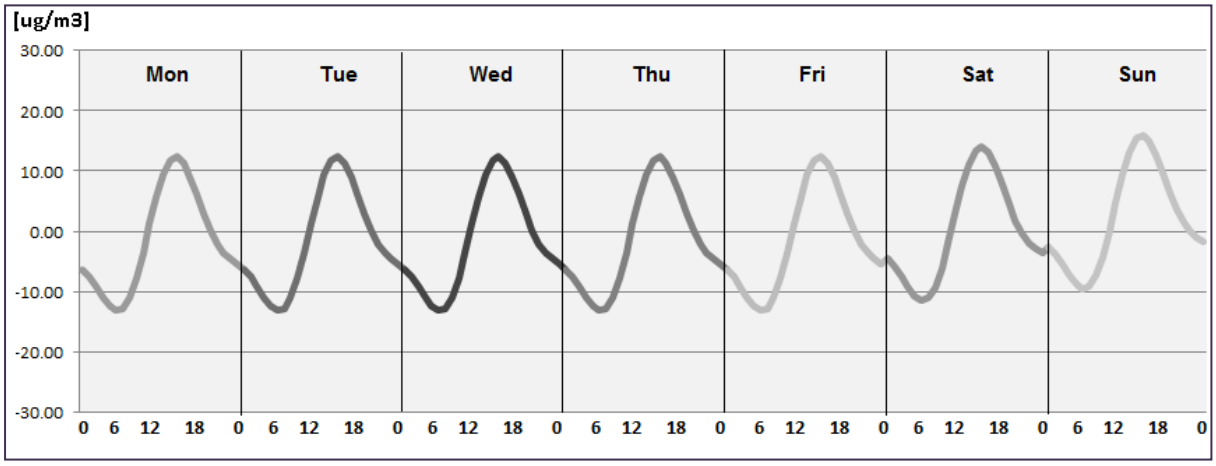


Figure 6.18 Weekly and diurnal cycle of modelled O_3 in site type 1

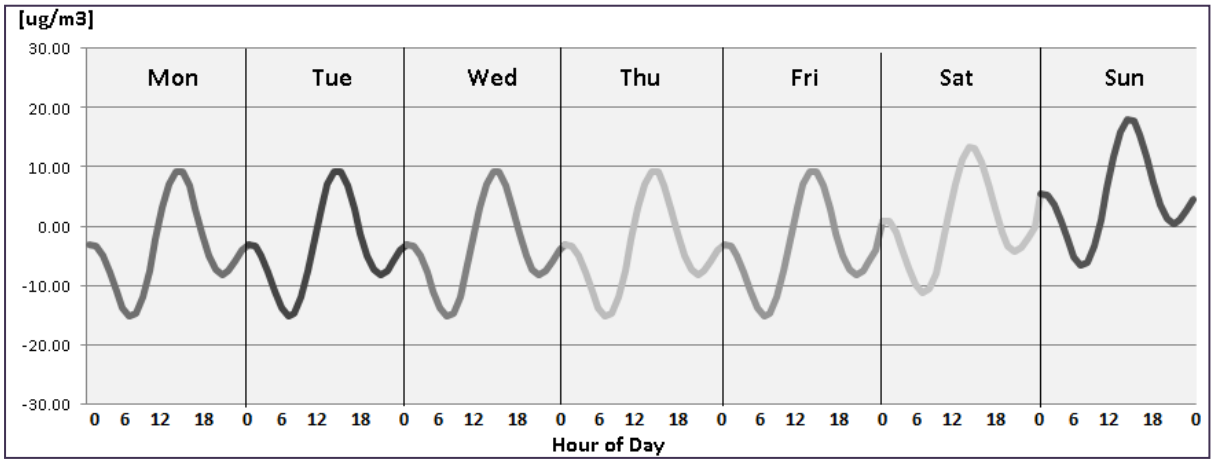


Figure 6.19 Weekly and diurnal cycle of modelled O_3 in site type 7

Figure 6.20 shows the relationship between predicted and observed concentrations (using the average of all sites in site type 1) for the first 2000 hours of the study period. In general, predicted concentrations track the measured values relatively closely, though there are periods (e.g. between hours 1552 and 1654) when the two lines diverge. The relatively long duration of these periods (a week or more) suggests that they may be associated with weather-related events unrelated to the systematic variation being modelled here. Some of the modelling error may also relate to short-term variations in emissions.

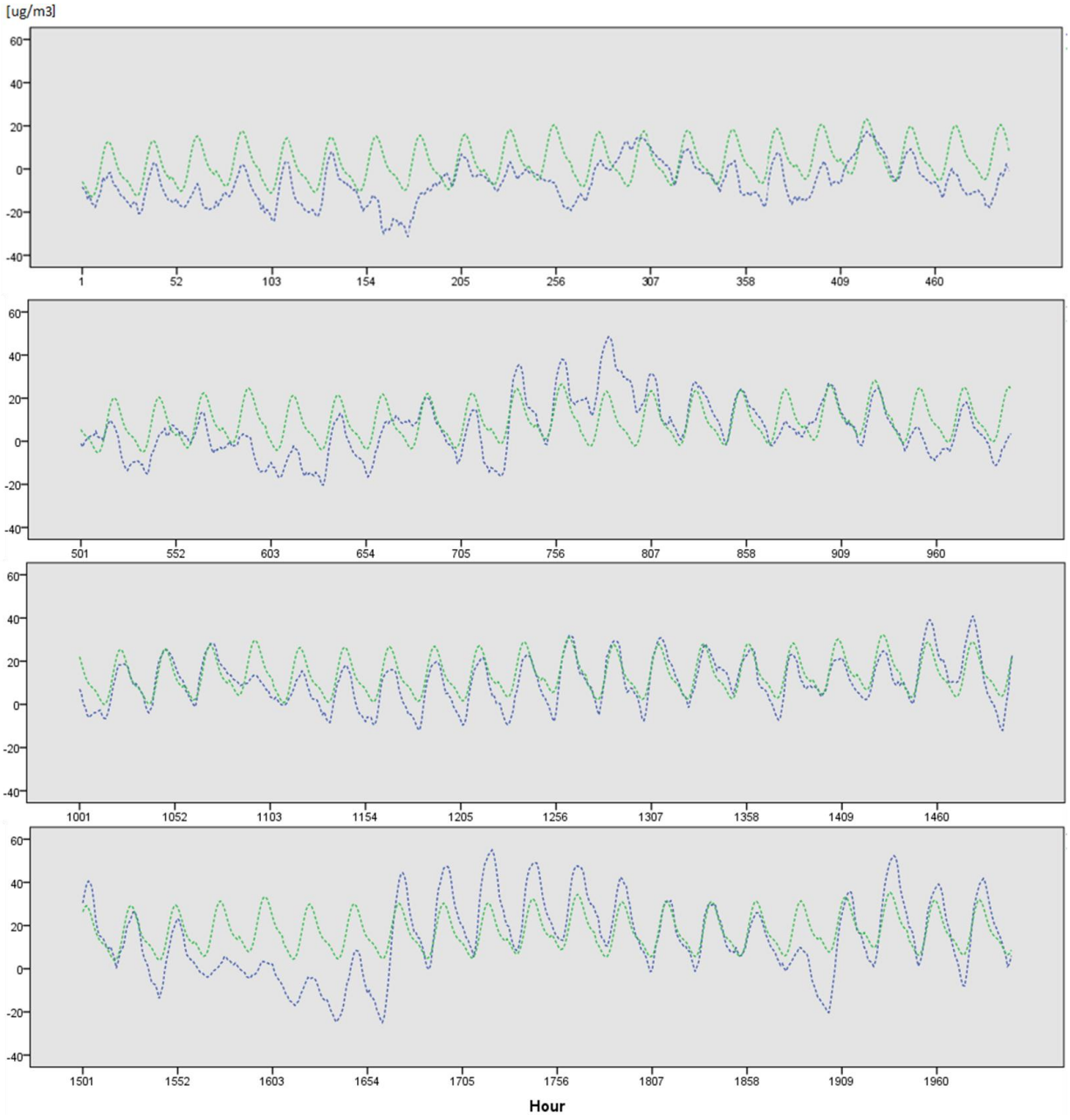


Figure 6.20 Predicted (green) and observed (blue) hourly O_3 concentrations (deviation from mean, $\mu g/m^3$) for the first 2000 hours, averaged across all sites in site type 1

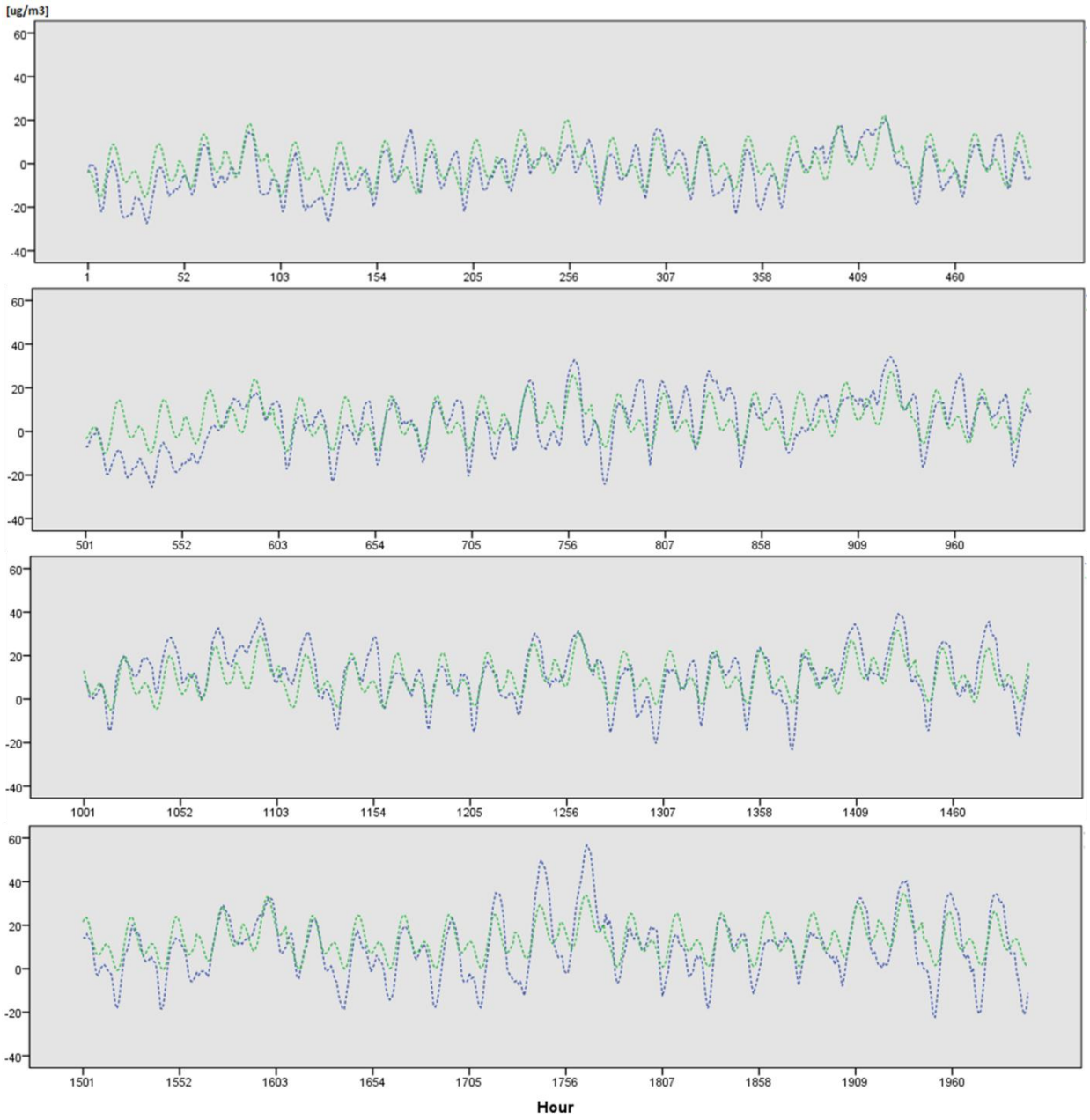


Figure 6.21 Predicted (green) and observed (blue) hourly O_3 concentrations (deviation from mean, μg/m^3) for the first 2000 hours averaged across all sites in site type 7

Figure 6.21 shows the hourly concentration (deviation from mean) for predicted and observed concentrations (using the average of all sites) for site type 7, for the first 2000 hours of the study period. Again, predicted concentrations track the measured values relatively closely, but at times the lines diverge (e.g. between hours 501-552), or the amplitude of variations in the predicted concentration is small compared to that of the observed (e.g. between hours 1705-1756 and 1909-

2000). Over-estimation occurs, for example, during the March period, especially hours 1-51 (week 1) and 501-552 (4th week of March), which broadly coincide with the Easter school holiday when there is reduced road traffic.

6.5.2 Results for all site types

Table 6.3 summarises the temporal models and the measures of goodness of fit to both the training and the validation data for each site type. The adjusted R^2 ranges from 0.27 to 0.58 and the RMSE between 20.5 and 26.8 $\mu g/m^3$. Applying the model to the independent validation data produces almost identical results, indicating that the model is stable and not over-fitted to the training data.

Table 6.4 shows the percentage of the overall temporal variability explained by these functions (i.e. that can be defined as systematic), and the proportion of this systematic variability represented by the seasonal, hebdomadal and diurnal functions.

Systematic variability accounts for 28% to 58% of the temporal variability (mean = 42%). Most of this is attributable to seasonal effects, which account for 63% of the systematic variation and consistently exceeds that of diurnal or weekly variability. Diurnal variability, however, is also substantial, accounting for 36% of the systematic variation. The weekly effect is small, and only accounts for 1.4% of the systematic variability on average.

6.5.2.1 Seasonal patterns

Seasonal variability is thus seen as the main source of variation in the data. This is represented in most site types by one seasonal time function that has one wave, peaking between spring and summer seasons (Table 6.3). The site types nevertheless differ to some extent both in the exact timing of the peak (from May to August) and in their coefficient (i.e. amplitude). Differences in timing seem to reflect differences in the regional environment of the different site types. In maritime areas, for example, the peak tends to be in spring (S1: maximum in May approximately), as in site type 8; in both southern and inlands site types, in contrast, the maximum is shifted to the summer (S3: maximum at July approximately) as in site types 11 (Southern urban uplands), and 4 (Urban inland moderately sheltered). Other site types have a prolonged maximum between spring and summer, with the peak in June or thereabouts (S2: maximum at June).

Several site types (2, 12, and 13) show more complex seasonal patterns, and to model these second post hoc seasonal function (PHF) was incorporated. In the case of site type 12 (Forested mountain) the pattern is characterised by a strong and broad spring-summer maximum with the peak in spring (Figure 6.22). Site type 13 represents sheltered lands in southern Europe. This shows a strong peak in summer, probably reflecting the marked differences in solar radiation and temperature between summer and winter in these areas, amplified perhaps in this case by effects of the accumulation of stagnant air during hot, dry periods of the year. Site type 2, which has a more coastal distribution, has a wider and less marked maximum period, between spring and summer, indicating the damping effect of the maritime climate.

Several consistent trends can be seen in the distribution of these seasonal effects, by examining the percentage of systematic variation in Table 6.4 and the average environmental characteristics from Table 4.10 for each site types (Appendix B, Section IX). The results show that, across the thirteen site types, the amount of temporal variability in seasonal O_3 concentrations increases significantly with increasing altitude (R=0.59) and topex (R=0.72), and with non-agricultural land cover[24] (R=-0.57). The seasonal variation thus tends to be greater in site types with higher altitude, most notably in site types 1 and 12 where altitude is higher than >500m, topex is greater than 20m and non-agricultural land typically makes up more than 60% of the land area within 1 km. In general terms, also, the amplitude of the seasonal patterns tends to increase from northern to southern Europe. These findings are in line with findings of other studies, as reported in section 6.1 and imply that the models are picking up the general seasonal patterns in O_3 concentrations.

6.5.2.2 Hebdomadal and diurnal variability

Figures 6.23 - 6.24 show line graphs of the modelled systematic variation in the hourly concentration for one full week (from Monday to Sunday). Several site types (e.g. 3 and 6) show a simple pattern with a single, strong afternoon peak. Site type 12 shows a broadly similar pattern, but with much reduced amplitude. All these site types are rural in character, and lack local emission sources of precursors (e.g. from traffic or industry). These patterns thus seem to represent sites in which there is a relatively consistent pattern of O_3 production by photochemical reaction during the afternoon, followed by dispersion during the night and morning.

[24] 100 - (highdr_1000+Lowdr_1000+IND/COM_1000+Agr_1000)

In the majority of site types, however, there is evidence for a smaller, secondary peak (or at least, an inflexion in the curve) during the night-time (typically at around mid-night). To explore this, the residuals for the night period (between 10:00 pm and 3:00 am) were examined by plotting a scatterplot against the observed O_3 concentrations, coding the data points by night and day. The result showed no difference in the two patterns, indicates that small peak at night represents a real feature of the data. However, this pattern is most marked in more urbanised site-types, such as site type 7 (see Figure 6.19), site type 11 (Southern urban upland) and site type 8 (maritime urban moderately sheltered). In urban areas, the physical processes that might generate such a pattern are well-established: following a period of dispersion and scavenging after the afternoon peak, O_3 tends to build up again as a result of dispersion from neighbouring rural areas.

The hebdomonal patterns also differ somewhat from one site type to another. Site type 7 stands out as having the greatest degree of hebdomonal variation, making up 7% of the total systematic variation in O_3 concentrations. Across the thirteen site types (Appendix B, Section IX), the magnitude of the difference between O_3 concentrations on weekdays and those on Sunday increases significantly with increasing urban area (R=0.90) and local road density (R=0.81), and with reductions in rural land cover (R=-0.84), altitude (R=-0.77), and Topex (R=-0.54). The weekend increment thus tends to be strongest more urbanised site types, with heavy traffic – notably in site types 7 and 9, where the weekday to Sunday difference is over 8ug/m^3. The smallest effects are seen in the more rural, high altitude and less exposed sites: in site type 12, for example, where urban land typically makes up ca. 5% of the land area within 1 km of the monitoring sites, the weekday to Sunday difference is less than 1ug/m^3. These differences are in line with theory and suggest that the models are picking up genuine hebdomonal patterns in O_3 concentrations.

Table 6.3 The temporal model for each site type, showing the coefficients and statistics of goodness fit for training and validation datasets

Site type	Environmental characteristics	Equation	Training (R^2,RMSE)	Validation (R^2,RMSE)
1	Forest hill-lands	0.92+(23.98*S2)+(11.12*D16)+(3.78*D2_14)+(-1.69*weekday)+(1.92*Sunday)+(3.65*PHF1)	0.38, 24.4	0.37,23.6
2	Sunny mixed use	1.17+(22.94*S2)+(12.87*PM5)+(5.30* D2_14)+(-2.03*weekday)+(1.77*Sunday)+(3.32*PHF2)	0.41, 23.2	0.44,23.2
3	Mixed use moderately sheltered	1.04+(22.73*S2)+(-2.00*weekday)+(2.04*Sunday)+(20.68*D15)+(5.61*D2_14)	0.44, 24.9	0.42,24.5
4	Urban inland moderately sheltered	1.82+(27.19*S3)+(-3.39*weekday)+(3.55*Sunday)+(17.61*D16)+(7.25*D2_14)	0.47, 24.9	0.47, 24.9
5	Urban inland	1.87+(22.34*S2)+(16.54*D16)+(5.57*D2_14)+(-3.27*weekday)+(2.57*Sunday)	0.42, 23.9	0.40,23.9
6	Sunny mixed use strongly sheltered	1.67+(22.38*S2)+(25.23*D15)+(6.68*D2_14)+(-2.94*weekday)+(2.62*Sunday)	0.48, 25.6	0.45, 26.3
7	Heavily trafficked urban	2.26+(15.43*S2)+(-4.05*weekday)+(4.58* Sunday)+(7.29*D16)+(7.02* D2_14)	0.30, 20.5	0.30 ,20.1
8	Maritime urban moderately sheltered	2.19+(15.93*S1)+(-3.78*weekday)+(3.37*Sunday)+(13.83*D15)+(6.09*D2_14)	0.31, 23.6	0.30,23.6
9	Northern Urban	2.35+ (20.67*S2) + (-4.29*weekday)+(3.58*Sunday)+(15.43*D16)+(5.49*D2_14)	0.37, 24.6	0.38,25.1
10	Inland populated strongly sheltered	1.33+(27.11*S2)+(-2.58*weekday)+(2.67*Sunday)+(22.28*D16)+(6.37*D3_15)	0.48, 26.5	0.49,26.5
11	Southern urban lands	2.317+(23.97*S3)+(-3.73*weekday)+(2.46*Sunday)+(19.62*D16)+(10.07*D2_14)	0.51, 22.9	0.50, 22.6
12	Forest mountain	0.55+(16.50*S1)+(6.20*D17)+(2.76*PHF3)+(-0.55*weekday)+(0.35*Sunday)	0.27, 21.3	0.27, 22.2
13	Southern populated strongly sheltered	2.08+(36.69*S3)+(20.41*D17)+(7.69*D4_16)+(5.26*PHF4)+(-2.69*weekday)+(3.16*Sunday)	0.58, 26.8	0.55,26.0

Post-hoc time Functions: (PHF1=SIN(4*3.142*(DoY-100)/365)); (PHF2=COS(4.5*3.142*(DoY-70)/365));(PHF3=COS(4.5*3.142*(DoY-80)/365));(PHF4=SIN(3.5*3.142*(DoY+60)/365)).

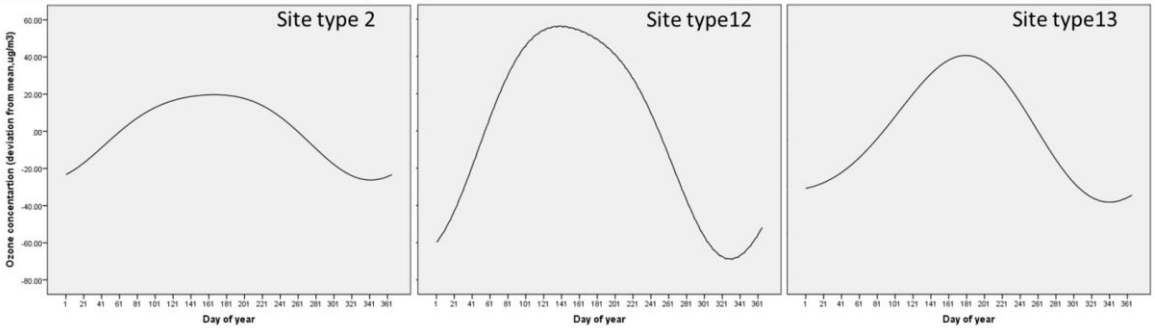


Figure 6.22 Modelled seasonal variability in site types 2, 12, and 13

Table 6.4 Systematic variation as a percentage of total temporal variability in O_3 concentration explained by time functions for different time periods and the dominant characteristics for each site type

Sitetype / Temporal variation	1	2	3	4	5	6	7	8	9	10	11	12	13	Average
Systematic variability	**38**	**41**	**44**	**47**	**42**	**48**	**30**	**31**	**37**	**48**	**51**	**27**	**58**	**42**
Seasonal [B]	79	71	52	66	61	41	67	52	60	56	53	85	76	**63**
Hebdomonal [B]	0	0	0	2	0.0	0	7	3	3	2	2	0	0	**1**
Diurnal [B]	21	29	48	32	38	58	27	45	38	42	45	15	24	**36**
Character of site type [C]														
Urban (%)	18	47	39	75	62	42	80	60	63	57	67	5	61	
Rural (%)	82	53	61	25	38	58	20	40	37	43	33	95	39	
High (>500 metres)	X											X		
Low (<200 metres)					X			X	X					
Maritime (<150km)		X					X	X					X	
Inland (>250 km)	X			X						X				
North	X						X		X					
South	X						X				X		X	

A. Systematic variability as a percentage of the total temporal variability
B. Percentage of the systematic variability
C. Most dominant characteristics in the site type

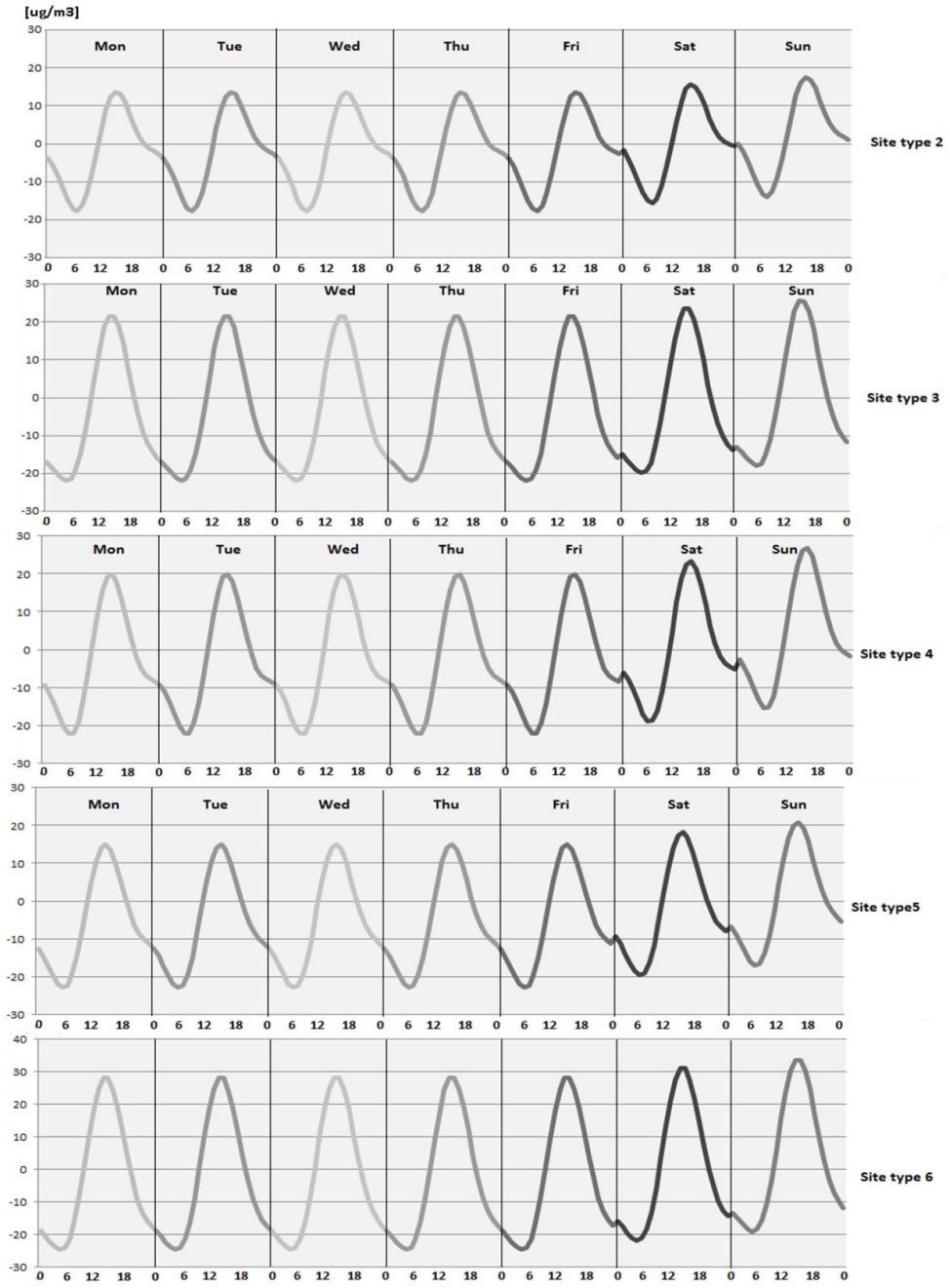


Figure 6.23 Line graph of modelled hourly variation during 7 days, from Monday to Sunday, for site types 2 – 6

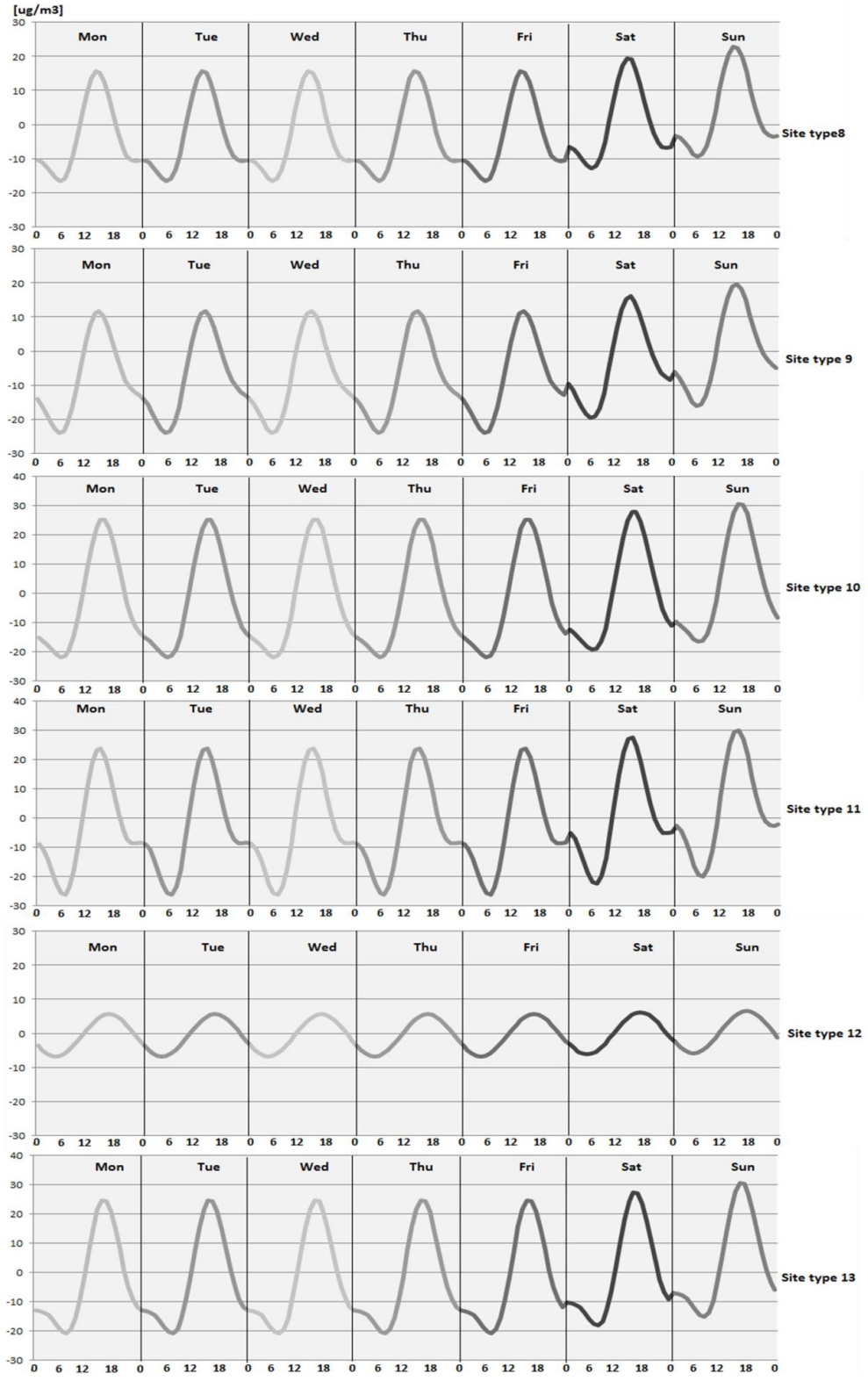


Figure 6.24 Line graph of hourly variation during 7 days, from Monday to Sunday, for site types 8-13

6.6 Component of variability

As stated earlier, the modelling approach adopted here recognises three main components of temporal variation: 1) systematic variation; 2) unsystematic, temporally correlated variation; 3) random variation (noise). In this chapter, Fourier analysis has been used to generate time functions to simulate the systematic temporal variability of O_3 concentrations over different site types across Western Europe. Systematic variability operates over three different time scales: seasonal, hebdomadal and diurnal.

The resulting models show generally similar patterns of pollution across all site types, with a clear afternoon peak of concentrations that reflects a more-or-less universal accumulation over the day as photochemical activity builds up. Detailed differences are, however, seen between the site types, especially in the amplitude and timing of the afternoon peak, the width of the peak and, in many cases in the occurrence of a smaller secondary night-time peak.

Systematic variability is seen to make up less than half of the total temporal variability in O_3 concentrations – typically about 42%. The remainder is either random temporally correlated variability or noise.

The relative importance of the three different components of variability (i.e. time scales), broadly reflect the characteristics of the site types. Most of the temporal variability is associated with the seasonal pattern (winter/summer), which accounts for 41%-85% of the systematic variability (mean = 63%). Diurnal patterns account for most of the remainder (15-58%, mean = 36%). While the hebdomadal effect is negligible, accounting for no more than 1% of the total systematic variability, in more urbanised sites the weekend increment may amount to 8ug/m^3 or more. Notably, these patterns broadly reflect the results found previously (Chapter 4, Table 4.3) using VCA. This showed that temporal variability explained about 28% of overall variability (including spatial) and within this, seasonal variability was dominant, accounting for about 65% of the temporal variability and diurnal variability about 34%. This suggests that the Fourier models are successfully capturing the majority of the systematic variability in the data.

The different components of the systematic variation in the Fourier models vary in their importance geographically. As a proportion of overall temporal variability (Appendix B, Section IX), systematic variation increases with decreasing topex (R =-0.663 across the 13 site types), implying that in topographically exposed areas there is more random variability, probably due to short-term

variations in weather conditions and the lack of long periods of O_3 accumulation as may occur in valleys. The relative importance of these different components of systematic variation likewise varies with topography and, to a lesser degree, with land cover. Seasonal variation, as a proportion of total systematic variability, increases both as topex increases (R = 0.72) and more weakly as altitude and the area of non-agricultural rural land increase (R=0.59 and 0.57 respectively). Systematic seasonal patterns are thus strongest in upland, exposed rural areas, where the extremes of temperature and photochemical activity are most marked.

The pattern of seasonal variability also differs between different site types. In general, most site types show a prolonged maximum between spring and summer seasons. More mountainous and maritime sites types, however, are characterised by an earlier peak in the spring, while in the forested hill-lands, the seasonal maximum tends to occur in summer.

Diurnal variability shows a less clear pattern (Appendix B, Section IX), but tends to increase as topex and altitude decline (R=-0.71 and -0.56 respectively). Diurnal variability is thus strongest in lowland, valley situations, where stagnant air can accumulate.

6.7 Summary

Similar to spatial modelling, the temporal modelling also attempts to model the three elements of variability, only in this case in the temporal dimension. The Fourier models developed here were generated semi-deterministically – by designing a priori functions to reflect the expected patterns of systematic variability in the data. This is justified given that most of the temporal variation in O_3 concentrations is related to systematic variations in temperature and solar radiation and, perhaps to a lesser extent, in human behaviour.

Using Fourier analysis to generate the systematic variation based on knowledge from previous studies and theory has its advantages and disadvantages. An advantage is that it helps to ensure that the patterns are consistent with the environment for which the models are built (e.g. for urban or rural areas). This makes the extrapolation of the models to other unmonitored locations with the same underlying characteristics safer. By the same token, it is easier to interpret and explain the processes behind the patterns observed, and to use such interpretations as a check upon the veracity of the models. Modelling the systematic variation statistically (for example using ARIMA analysis or polynomial functions – ARIMA) is likely to result in models that better fit the observed

data, but without the same assurance that the patterns are both physically plausible and generalisable to other study areas (Kumar and Jain, 2010). The major disadvantage of using Fourier analysis is that the time functions need to be able to match both the shape of the patterns of O_3 variation, and the timing of the variations. Building these into the models requires a sound understanding of O_3 processes and the way these change in different environmental contexts. Even with this, surprises may occur and it may be necessary to incorporate additional post hoc functions to describe the patterns effectively. Inevitably judgements have to be made at this point about the plausibility of these functions. As with all such models that take no account of information that might be available in other covariates, the Fourier models can only represent the systematic variation in O_3 concentrations. Modelling the unsystematic variability with any degree of reliability requires the incorporation of additional data on the factors that contribute to such variation.

Attempts were made to model some of the remaining variation by incorporating additional trigonometric functions, but these did not always improve the model. This suggests that the remaining variation is largely non-systematic. Some of this residual variation nevertheless shows temporal patterns, often in the form of episodic behaviour – i.e. periods of high or low O_3 concentrations which are under- or over-estimated by the Fourier models. These might be expected to be associated with weather events. Thus, in the next chapter (Chapter 7), attempts are made to model this residual variation using time-varying meteorological data by applying the full model at two different spatial scales: one within a country and another within a city having a different weather regime.

Part3: Model Application

7 Space-time O_3 model

Previous chapters have described the development of the temporal and spatial components of the model. Finally, the two models need to be combined to provide a space-time model. This chapter explains how the weighted time function models (TM 1 to TM13) were linked to the spatial model, to generate the space-time models and its application to sites across Europe and in two case studies areas (NL and Rome).

The space time models were developed in two stages which are referred to as the base space-time model and the full space-time model (Figure 7.1).

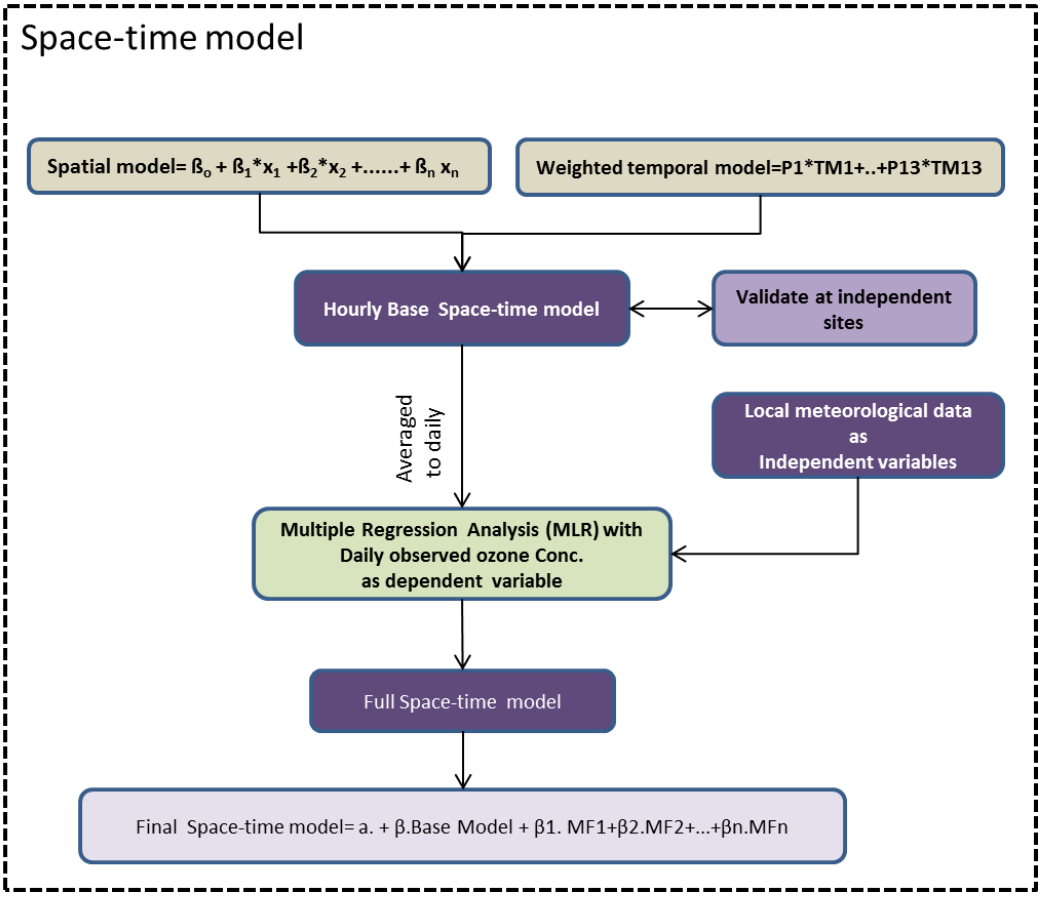


Figure 7.1 Steps in building the space-time model (Base and Full models)

Construction of The base space-time model (Section 7.1) was the first step in combining the spatial and temporal models and was achieved in two steps as follows (see also Figure 7.1):

1. Site type membership probabilities from MLOR analysis were used as weighting factors for each time model to estimate the temporal variability in O_3 concentrations at each location.
2. These estimates were then combined with estimates of the long term mean concentration, derived from the spatial model, using two approaches: either simple addition or by calibration with measured concentrations in a multiple regression analysis.

This model was used to generate hourly and daily concentration estimates for the monitoring sites across Western Europe, and hourly estimates validated using an independent subset (discussed in section 7.1).

Subsequently, attempts were made to improve these estimates by including local meteorological information to represent the time-varying, non-systematic weather-related influences on O_3 concentrations. This improved model is referred to as the full space-time model. As it was not possible to obtain hourly or daily meteorological data for the whole study area, the full space-time was applied in only two areas:

1. The Netherlands (Section 7.2), where the model was used to estimate O_3concentrations for a 100m grid covering the whole country and validated using two approaches.
2. The city of Rome (7.3), where the model was used to estimate concentrations for point locations representing the homes of cohort participants.

These areas were selected with the additional aim of providing daily estimates of O_3 exposures for birth outcome cohorts in the ESCAPE project.

The performance of the full space time model is further investigated by completing a short exposure assessment study (Section 7.4) for both case study areas.

7.1 Base space-time model

As mentioned previously, the spatial model (derived by LUR) gave the long-term mean concentration between March 2001 and February 2007, while the weighted time models provided estimates of the hourly variation around the long-term mean. The question, however, arises of how these two components of the modelling should be combined: are the two models, for example, simply

additive; or is a calibration with the European monitoring sites required to combine the two components?

Both possibilities were explored, to produce two different versions of what is referred to here as the base model (i.e. that using only the LUR model and the weighted time functions):

1- Additive hourly base model, by direct addition:

   Predicted O_3 = LUR+WTM hourly

2- Calibrated hourly base model, by calibration:

   Predicted O_3 = constant + (B_1*LUR) + (B_2*WTM hourly)

In developing the calibrated hourly model, regression analysis was done using the training monitoring sites, and the results validated by the validation dataset. For comparison, the additive hourly model was also tested against both data sets.

For the calibrated model, the following regression equation was derived.

Hourly predicted O_3 (B) = -0.0223 + (1.00*LUR) + (1.017*WTM)

An almost identical model was obtained when using the daily O_3 concentrations:

Daily predicted O_3 = - 0.259+ (1.006*LUR) + (1.003*WTF)

Information on the goodness of fit for the hourly models is shown in Table 7.1.

Table 7.1 Summary of the validation results for both additive and calibrated hourly base models

The model	dataset	R	Adj.R^2	RMSE
Additive model	**Training dataset**	**0.69**	**0.47**	**25.29**
	Validation dataset	**0.68**	**0.46**	**25.38**
Calibrated model	**Training dataset**	**0.69**	**0.47**	**25.30**
	Validation dataset	**0.68**	**0.46**	**25.38**

As these results indicate, there is almost no difference either in the form of the calibrated and additive hourly models, nor in their performance when compared with the training and validation datasets.

Descriptive statistics (min, max, and SD) for predicted concentrations from both the additive and calibrated models were compared with observed concentrations at the monitoring sites (training dataset) (Table 7.2).

Table 7.2 Descriptive statistics for additive and calibrated model and observed concentrations

Variable	Min	Max	Mean	SD
Hourly observed O_3	0	470.00	49.80	32.23
Hourly predicted O_3 (additive)	0	137.55	49.60	22.80
Hourly predicted O_3 (calibrated)	0	136.84	49.58	23.01

As is to be expected, the results for the two methods are again more-or-less identical. Based on these results, it was therefore decided that the simple additive model could be used for subsequent analysis as this represented the more straightforward, and more logical, approach. Predicted hourly concentrations using the additive model were therefore aggregated to three different time periods - diurnally, weekly and monthly - and compared with the observed concentrations to explore the correlation between predicted and observed at different time scales. Results are summarised in Table 7.3.

Table 7.3 Performance statistics for additive base model: Pearson correlation, R2, and RMSE between observed and predicted concentration for different time scales

Variable	R	Adj.R^2	RMSE
Daily O_3 concentration	0.73	0.53	18.47
Weekly O_3 concentrations	0.81	0.65	14.28
Monthly O_3 concentrations	0.86	0.74	11.40

These results show that, as is to be expected, aggregating the concentrations to longer periods increases the proportion of explained variation in O_3 concentrations and reduces the RMSE.

The base model predicts only the systematic temporal variation in O_3 concentrations, which is repeated over the whole time period. Any non-systematic variation is left unexplained. The model thus generally fails to predict extreme values (high or low) which typically occur irregularly on the hourly and daily concentrations, as indicated in Figures 6.20 and 6.21. This is also reflected in the relatively high RMSE values in Table 7.3, especially when the averaging times are short.

These more irregular variations in O_3 concentrations are likely to be driven, in part, by episodes either in meteorological conditions (e.g. heat-waves, blocking antic-cyclonic conditions causing

prolonged inversions) or in emissions (e.g. holiday periods, strikes). It is important to recognise, however, that these two influences do not operate wholly independently; changes in emissions, for example, often occur as a direct response to changes in weather conditions (e.g. increased combustion for heating during cold-spells, or air-conditioning during heat-waves). Time-varying data on emissions are not readily available, but daily meteorological data do exist for a dense network of stations across Europe (although access to these is sometimes limited for different countries). As mentioned, for NL and Rome, the meteorologically-driven effects (both direct and indirect) were modelled for two reasons:

1. First, to illustrate how the full model can be further developed and to determine whether adding meteorological factors will represent the non-systematic temporally-correlated variability effectively, and provide estimates of the extreme values that the base model tends to miss.

2. Second to estimate the daily concentrations for these study areas in order to provide exposure estimates for use in the ESCAPE project.

The following section explains how this was done, on an exploratory basis, for these two study areas.

7.2 Full space-time model in the Netherland (case study 1)

The Netherlands is a small country in North-West Europe and shares its border with Belgium in the south and Germany in the east, and is bounded in the west by the North Sea. There are twelve provinces in the country, and the capital is Amsterdam. Geographically the Netherlands covers an area of 41,543km^2 of which 33,883 km^2 is land and the rest water. It is inhabited by 16,731,092 people according to the latest estimate of Statistics Netherlands[25]. As the name (which means "the low country") indicates, the topography is relatively flat and low-lying, with 25% of the land below sea level, and 50% less than one metre above the sea. The Netherlands is characterized by a maritime climate, with a narrow annual range of temperature and precipitation throughout the year.

There are two prospective birth cohort studies being conducted in the Netherlands: PIAMA (Prevention and Incidence of Asthma and Mite Allergy) and ABCD (Amsterdam Born Children and their Developments). PIAMA recruited 10,819 pregnant women between March 1996 and May 1997, located all over the Netherlands, though clustered in the north, west and centre. One of its aims is to

[25] http://www.cbs.nl/en-GB/menu/themas/bevolking/cijfers/extra/bevolkingsteller.html

evaluate the natural history of asthma and allergy in association with many factors, including air pollution (Brunekreef et al., 2002). The ABCD cohort is located in Amsterdam and recruited 12,682 pregnant women between January 2003 and March 2004, to examine the relationship between maternal lifestyle and birth outcome (Van Eijsden et al., 2006). Neither study has evaluated the association with ambient O_3, although both have previously analysed other pollutants (Pereira et al., 2012, Gehring et al., 2011a, Gehring et al., 2011b). Both, however, are included in the recently funded ESCAPE project and this case study was carried out to provide O_3 exposure estimates for the two cohorts.

7.2.1 Methodology

Selection of meteorological data

According to the information from the Dutch Royal Meteorological Institute[26] there are thirty six meteorological stations measuring the required meteorological factors (temperature, wind speed, sunshine and precipitation) at a daily level in the Netherlands. These variables are those most commonly included as variables for modelling O_3, as mentioned in Chapter 3.

The meteorological data varies very little between stations in the Netherlands, partly because of its relatively flat topography and maritime climate. For example, correlations were calculated between daily values at pairs of stations in the east of the country (where topography is more variable and the climate more continental), using a random sample of 30% of the data. Between-site correlations varied between 0.93 and 0.99 for three of the variables used here; for the fourth (total precipitation) the range was between 0.75 and 0.80. To further assess the possible effects of site-choice on the modelled estimates, a sensitivity analysis was also undertaken. In which precipitation data from different sites was applied. Results are presented in section 7.2.2. Based on these results, it was considered valid to use only one meteorological station for each area of the country. For this analysis, therefore, meteorological sites were incorporated into the analysis as follows:

- One meteorological station was selected in each of the north, south, east and west of the country.
- Stations were selected according to the following two criteria:

[26] Available in <www.knmi.nl>

- Each meteorology station had to be located in open space;
- Daily data had to be available during the full study period >75%.

- The country was then divided into four meteorological zones, centred on these stations, using Thiessen-polygons;
- O_3 monitoring sites were attributed to their relevant meteorological zone, and thus meteorological site, by point-in-polygon analysis.

Figure 7.2 shows the distribution of the meteorological stations in relation to the O_3 monitoring sites. Table 7.4 summarises the meteorological data for each of the four selected stations. It is evident from these data that variations in meteorology between the four stations is small, reflecting the small size and simple topography of the country.

Table 7.4 Descriptive statistics for the four meteorological stations in the Netherlands

Meteorological stations	**N***	**Wind speed (m/s)**			**Temperature (°C)**			**Sun duration (hour)**			**Total precipitation (mm)**		
Statistical measures	2191	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Schiphol	2191	.9	15.5	4.91	-6.1	26.7	10.95	0	15.5	4.81	0	56.7	2.43
Eelde	2191	.8	13.3	4.17	-9.6	25.5	10.05	0	15.5	4.62	0	51.3	2.34
Twenthe	2191	.5	10.8	3.45	-10.7	26.5	10.25	0	15.5	4.67	0	45.0	2.15
Gilze-Rijen	2191	.8	10.7	3.62	-7.9	28.0	10.82	0	15.5	4.71	0	53.6	2.25

*N: All data measured in the daily basis and available for 2191 days

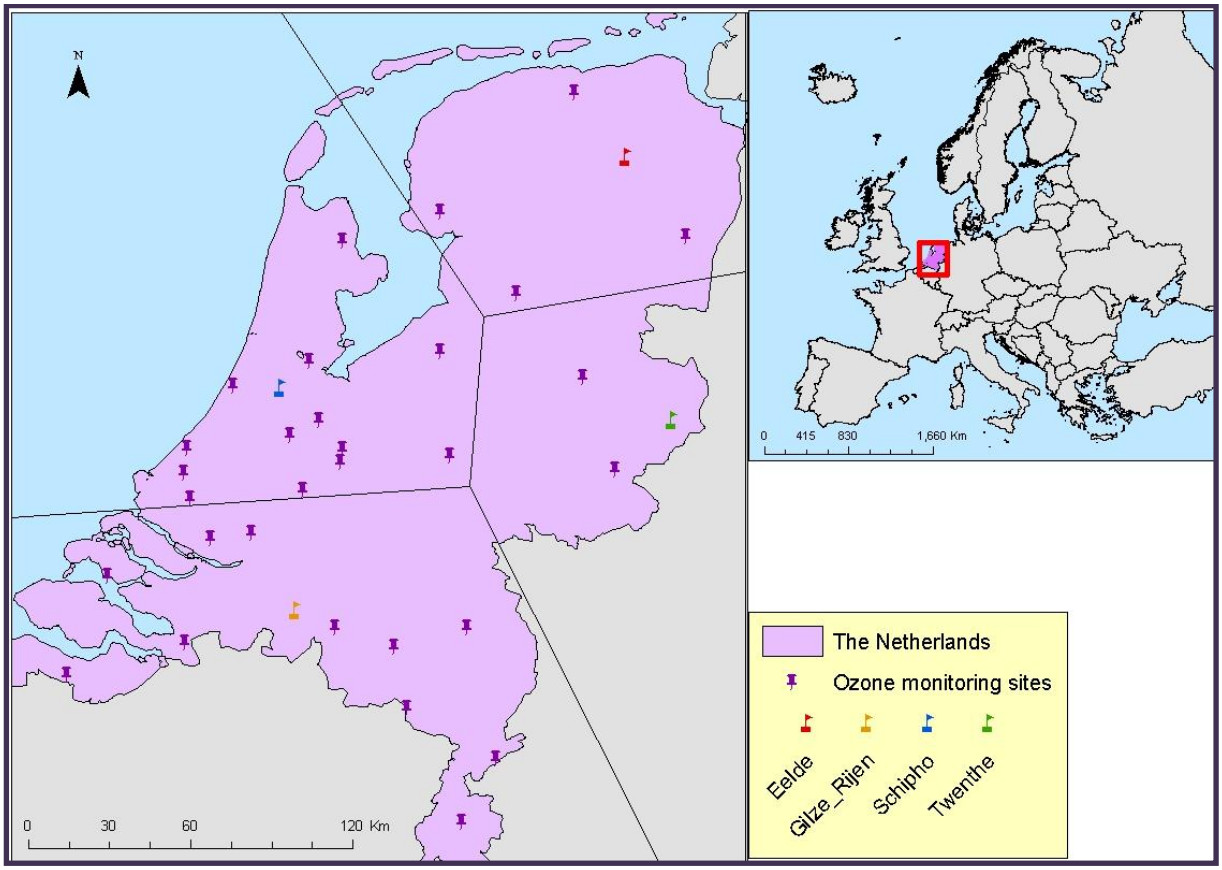


Figure 7.2 Map of the Netherlands showing the locations of the four meteorological stations (Eelde, Gilze, Scipho, and Twenthe) and thirty O_3 monitoring stations (purple pins)

Constructions of the full space-time model

The next step was to calculate the daily O_3 concentrations for each grid cell in the country. This was done by using the base model, which combines the spatial model (i.e. LUR) and the weighted daily time model (daily WTM), together with local meteorological factors, to develop a full space-time model. The model was built by simple stepwise multiple regression analysis, with the condition that each meteorological variable had to increase R^2 by 1% with a VIF<5. Daily observed O_3 concentrations from the thirty monitoring sites from both datasets (training and validation) provided the dependent variable. The analysis was done using all sites, because there were an insufficient number of sites to retain a separate validation dataset. However, the previous validation of the base model (Table 7.1) shows that the model performance was stable and this is not expected to bias the results. The LUR estimates, daily weighted time function models estimates, and daily unlogged values for the meteorological factors (listed in Table 7.5) were included as the independent variables

(predictors). From the regression analysis, coefficients were thus derived for each of these components in the full space-time model:

Predicted daily O_3= a + β. (Base model) + (β_1.MF_1+ β_2.MF_2 + β_3.MF_3 + β_4.MF_4)

where a is the constant, β is the regression coefficient and MF is the meteorological factors.

Nevertheless, to ensure the stability of the full model and assess the internal consistency of model results, it was tested using a leave-one-out cross-validation (LOOCV). This approach involves using a single site from the 30 sites as the validation data, and the remaining 29 sites as the training data. The validation site is then replaced by a different site, and the analysis repeated. This is done 30 times such that each observation in the sample is used once as the validation data. Then the 30 predicted values were regressed against the observed concentrations measured at the 30 monitoring sites and the regression statistics calculated (R^2 and RMSE). To provide further validation of the model, and the scope to apply it to another area not used in model development, it was also applied to 34 monitoring sites in the neighbouring country of Belgium. For this analysis meteorological data were obtained from the meteorological station located at the southern boundary of the Netherlands, on the assumption that it will represent more or less the same weather condition. Also, if the model works well in Belgium using the NL meteorological station, this would indicate that using meteorological data from a station located closer to the monitoring sites would improve the performance of the model.

After obtaining the coefficients for each variable in the regression equation, the probability of site-type membership at each 100m grid location had to be determined. This was done by applying the MLOR equation (described in section 4.2.4) to each grid centroid, using RASTER/MATH in ArcMap (Figure 7.3). Next, the thirteen time function models (TM) were weighted according to the probability of group membership for each grid cell, as specified in Figure 7.1, to provide the final weighted function time model:

WTM = P1.TM1 + P2. TM2 +.... + P13.TM13 **Equation 7-1**

The long-term mean O_3 concentration was then extracted from the LUR analysis of the whole study area using the EXTRACT VALUE TO POINT in ArcGIS (Figure 7.4).

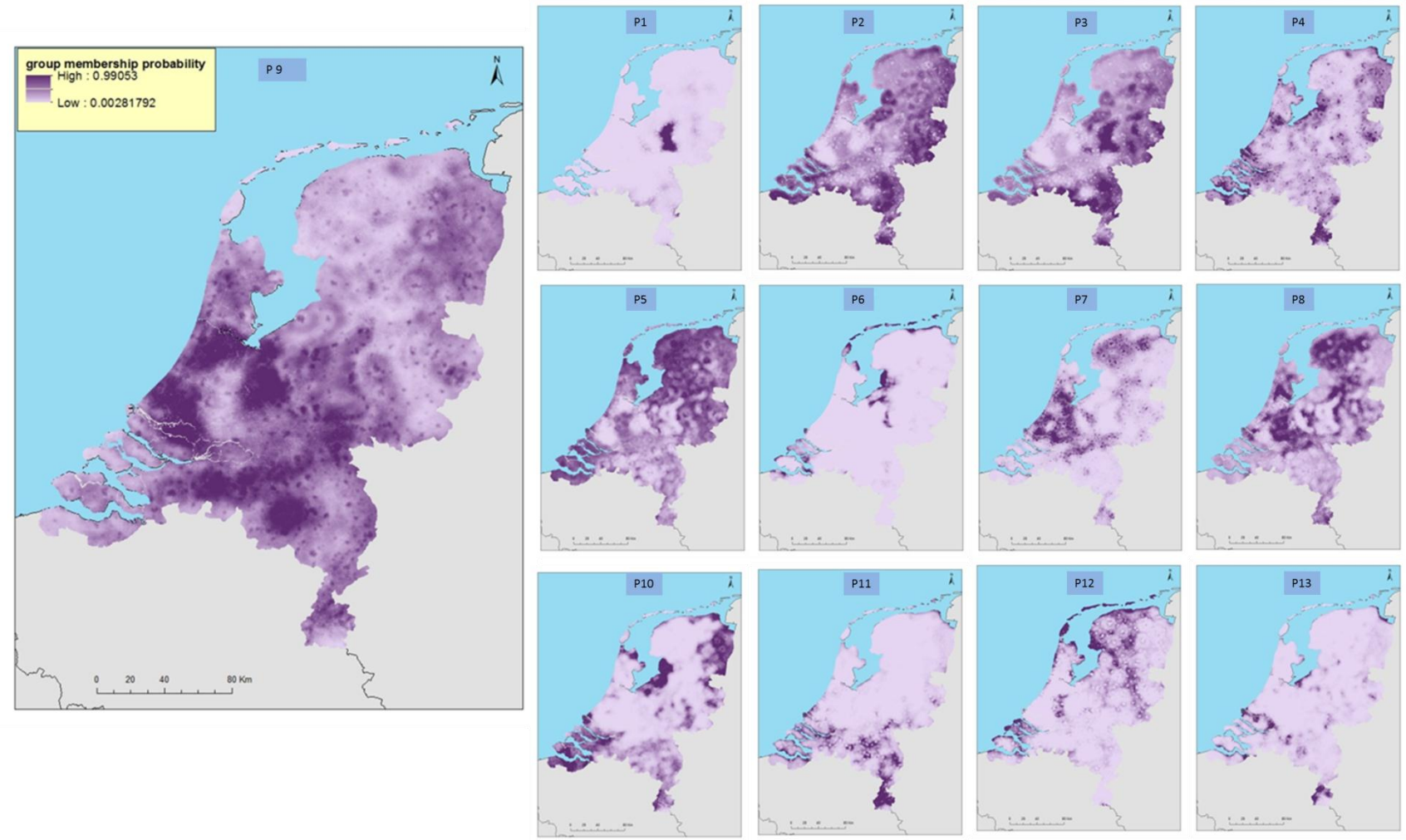


Figure 7.3 Site type membership probability (P1 to P13) for a 100 metre grid across the Netherlands

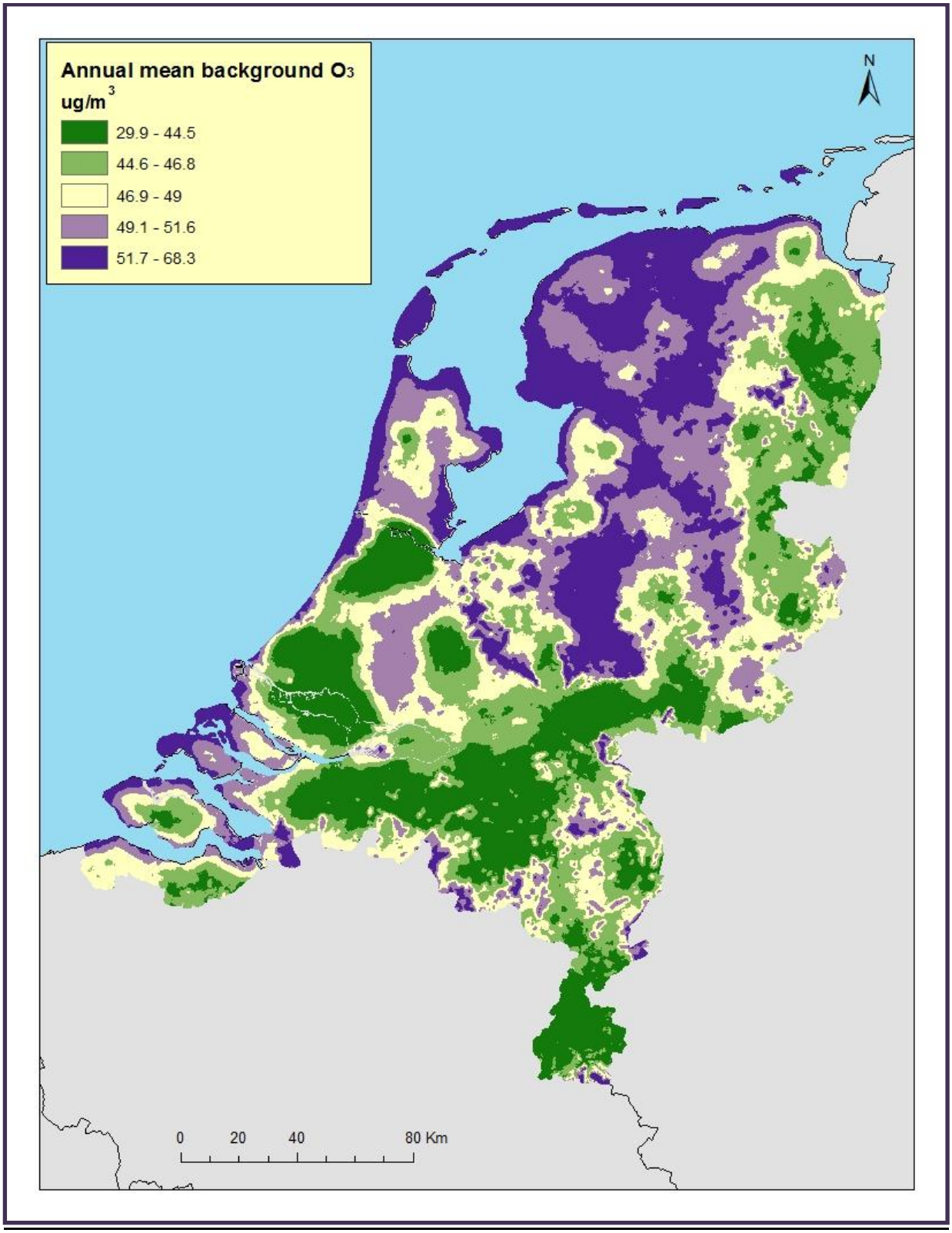


Figure 7.4 Long-term mean O_3 concentrations for the Netherlands estimated using the spatial model (LUR)

To develop a map of modelled concentrations, or to provide data for a dense network of receptors, the procedure above may have to be applied to a large number of target locations. For the Netherlands, at a 100 metre resolution, for example, calculations had to be done for 1.4 million grid cells. While this is possible in ArcGIS or a standard statistical package such as SPSS, it is computationally intensive, and for the Netherlands would have required two to three weeks of continuous processing time.

In order to speed up this aspect of the processing, therefore, the full model was run using Perl to calculate the weighted temporal component and then combine the output from the spatial model and weighted meteorological factors. The scripts are short text files and run the temporal component of the model in a matter of hours. The scripts were written and run by Margaret Douglass in the Department of Epidemiology and Biostatistics at Imperial College using calculation formulae provided by the author. As Appendix A, Section VII, shows, the scripts are short text files, designed solely to automate the computational procedures. They are also readily transportable to different computer platforms and operating systems.

7.2.2 Results and discussion

Results of the stepwise multiple regression analysis using all 30 sites are shown in Table 7.5. The meteorological variables each added ≥ 1% to adjusted R^2, except total precipitation (TP). This was therefore excluded from the final space-time model for the Netherlands. The final model for the daily predicted O_3 concentrations (C) is thus:

C =-18.4 + (0.77*Base model) + (3.1*WS) + (1.2*Sd) + (0.3*TEMP) **Equation 7-2**

where C is concentration, WS is daily windspeed, Sd is daily sun duration, and TEMP is the daily temperature value.

This model is consistent with expectations in that there are positive associations with all the meteorological factors. Using the absolute value of Beta (the standardised coefficient), it is apparent that daily O_3 concentration is affected greatly by both wind speed and sun duration. This is presumed to reflect the climate of the Netherlands. Wind speed, it seems, is acting as a proxy for vertical mixing and while the effect of sun duration is slightly weaker, it clearly represents the effect of solar radiation on photochemical activity. The contribution of temperature is smaller, partly

perhaps because variation is limited in the mild, maritime climate of the Netherlands (the correlation of temperature with sun duration is low, R = 0.38).

Table 7.5 Incremental statistics for the stepwise multiple regression analysis: summary of the final model

Model predictors	R	Adj. R^2	Change in R^2	Beta	RMSE	VIF
Base model	0.65	0.42		0.55	16.59	
Wind speed (WS)	0.70	0.48	0.06	0.29	15.66	1.10
Sun duration (Sd)	0.72	0.52	0.04	0.22	15.02	1.43
Temperature (TEMP)	0.73	0.53	0.01	0.10	14.96	1.54
Total precipitation*	0.73	0.53	0	0.03	14.94	1.18

*not included in the final model as does not contribute to R^2 by ≥1%

As Table 7.5 shows, the base model alone explains only 42% of the overall variability in the O_3 concentrations in the Netherlands sites. Inclusion of the meteorological variables gives a moderate improvement in model performance, with an increase in R^2 from 0.42 to 0.53, while RMSE falls slightly, from 16.6 to 14.9 $\mu g/m^3$. The improvement suggests that it is worthwhile including the meteorological factors in the model, though the full model still leaves 47% of the variation in O_3 concentrations unexplained.

As noted earlier, a sensitivity analysis was also carried out, to investigate the effects of using different meteorological data. Analysis was done by substituting the other two available stations for the eastern meteorological station used in the initial analysis. Performance of the full model did not change (R^2 and RMSE remained the same) and total precipitation still failed to increase the R^2 by more than 1% so was excluded from the model.

Table 7.6 Descriptive statistics for O_3 concentrations from the base and final models, compared to observed concentrations at thirty NL sites

Variable	Minimum	Maximum	Mean	RMSE
Observed O_3	0	144.73	39.97	
Base model	15.41	77.49	47.73	15.76
Full model	-1.83	80.93	39.88	14.94

As shown in Table 7.6, the full model predicted some concentrations below zero. This occurred only in 0.03% days (22 days with a negative value of the 65,730 in the full data set), all in the winter period (November and December). This might be for a number of reasons. One possibility is that the model is inadequately calibrated to the measured O_3 concentrations because the monitoring stations do not represent the full range of conditions in the country. Errors in both the input data and the parameterisation of the model are also, of course, inevitable. The negative values, however, might also represent a hidden reality. Under extreme conditions, the capacity for O_3 scavenging might exceed the O_3 supply in the atmosphere, thus creating potentially negative concentrations (though actual concentrations, of course, are bounded to zero). In this case, negative prediction values were substituted with zero.

As explained earlier, validation of the full model was also carried out, by applying a leave-one-out cross-validation. Results are shown in Table 7.7.

As the results show, the performance of the model varied somewhat across this analysis, with R^2 ranging from 0.44 to 0.65, and RMSE from 12.9 to 16.8 $\mu g/m^3$. When examined by region, there was little difference in the performance statistics. Sites located in the southern of the country showed a range in R^2 from 0.44 to 0.65; western sites from 0.46 to 0,61, northern sites from 0.45 to 0.58, and eastern sites were constant at 0.62. Overall, these results suggest that the performance of the model is stable and performance does not vary geographically.

For comparison, Table 7.8 summarises the performance of the daily base model and the full model at the thirty monitoring sites in the Netherlands, together with the results from the daily estimates derived from the LOOCV. It is clear that incorporating the meteorological factors made a moderate improvement in the predicted concentrations as R^2 increased, while the error decreased, at every site. Performance of the daily full model and the results from the LOOCV estimates were also similar, suggesting that there is no significant bias in the full model.

Table 7.7 Performance of the full models in the Netherlands sites using LOOCV

Left out site		**Prediction Model on N-1**							**Validation Results for left out site**	
ID of O_3 site	Location of Met site in the country	Model parameters: (Constant (Cons),base model (BM), Temperature(TEMP),windspeed(WS), sun duration(Sd) and Beta coefficient					R^2	RMSE	R^2	RMSE
		Con	BM	TEMP	WS	Sd				
NL00107	south	-18.46	0.77	0.24	3.11	1.16	0.53	14.96	.61	14.12
NL00131	south	-18.21	0.77	0.27	3.10	1.16	0.53	15.01	.58	13.51
NL00133	south	-19.06	0.78	0.25	3.12	1.17	0.53	14.99	.55	12.92
NL00227	south	-18.30	0.77	0.26	3.09	1.16	0.53	14.99	.63	13.46
NL00230	south	-18.39	0.77	0.26	3.09	1.16	0.53	15.02	.60	12.73
NL00235	south	-18.18	0.77	0.27	3.06	1.15	0.53	14.98	.56	13.64
NL00236	south	-18.48	0.77	0.26	3.10	1.15	0.53	15.00	.61	13.06
NL00301	south	-18.49	0.76	0.27	3.12	1.20	0.54	14.86	.44	16.57
NL00318	south	-18.60	0.77	0.27	3.10	1.18	0.53	14.93	.48	15.20
NL00404	west	-18.42	0.77	0.26	3.13	1.17	0.53	14.94	.52	15.46
NL00411	west	-18.38	0.77	0.27	3.13	1.17	0.53	14.96	.54	14.94
NL00433	west	-18.19	0.76	0.29	3.16	1.18	0.53	15.00	.55	13.03
NL00437	south	-18.47	0.77	0.29	3.01	1.18	0.53	14.98	.55	13.67
NL00441	south	-18.46	0.77	0.25	3.11	1.15	0.53	14.98	.65	12.96
NL00444	west	-18.35	0.77	0.28	3.09	1.17	0.53	14.86	.45	16.77
NL00520	west	-18.35	0.76	0.29	3.18	1.15	0.53	14.96	.53	14.53
NL00538	west	-18.61	0.77	0.29	3.12	0.17	0.53	14.92	.46	15.28
NL00620	west	-18.36	0.77	0.27	3.15	1.16	0.53	14.95	.61	13.51
NL00631	west	-18.51	0.77	0.28	3.14	1.17	0.53	14.97	.53	14.43
NL00633	west	-18.36	0.77	.27	3.14	1.16	0.53	14.95	.57	14.18
NL00636	west	-18.30	0.76	0.29	3.17	1.17	0.53	14.95	.55	13.75
NL00639	west	-18.13	0.76	0.30	3.16	1.17	0.53	14.98	.58	12.77
NL00641	west	-19.05	0.78	0.27	3.23	1.15	0.55	14.67	.46	13.79
NL00722	East	-18.25	0.77	0.28	3.09	1.14	0.53	15.00	.62	13.24
NL00738	west	-18.29	0.77	0.27	3.13	1.15	0.53	14.98	.59	14.50
NL00807	East	-18.17	0.76	0.28	3.10	1.14	0.53	14.97	.62	14.08
NL00818	North	-18.24	0.76	0.28	3.12	1.16	0.53	14.97	.58	14.39
NL00918	North	-18.18	0.76	0.29	3.12	1.17	0.53	14.94	.55	14.59
NL00929	North	-18.62	0.77	0.29	3.10	1.16	0.53	15.01	.55	12.89
NL00934	North	-18.57	0.76	0.31	3.11	1.17	0.53	14.92	.45	15.22

Table 7.8 Performance of the daily base, full models and the validated daily models in the Netherlands sites

	Daily Base model		Daily Full model		Validated daily models	
Monitoring site	Adj.R^2	RMSE	Adj.R^2	RMSE	Adj.R^2	RMSE
NL00107	.46	16.77	.62	14.03	.61	14.12
NL00131	.43	15.80	.60	13.41	.58	13.51
NL00133	.40	14.96	.56	12.88	.55	12.92
NL00227	.46	16.18	.63	13.45	.63	13.46
NL00230	.43	15.28	.60	12.74	.60	12.73
NL00235	.41	15.75	.56	13.68	.56	13.64
NL00236	.45	15.57	.61	13.15	.61	13.06
NL00301	.36	17.82	.45	15.50	.44	16.57
NL00318	.36	16.89	.48	15.19	.48	15.20
NL00404	.43	16.89	.52	15.42	.52	15.46
NL00411	.43	16.50	.54	14.93	.54	14.94
NL00433	.42	14.79	.56	12.99	.55	13.03
NL00437	.41	15.63	.55	13.65	.55	13.67
NL00441	.49	15.55	.65	12.91	.65	12.96
NL00444	.34	17.40	.45	16.77	.45	16.77
NL00520	.42	16.07	.53	14.52	.53	14.53
NL00538	.38	16.39	.46	14.87	.46	15.28
NL00620	.45	15.92	.61	13.42	.61	13.51
NL00631	.44	15.86	.53	14.43	.53	14.43
NL00633	.43	16.36	.57	14.15	.57	14.18
NL00636	.43	15.49	.55	13.80	.55	13.75
NL00639	.44	14.78	.59	12.73	.58	12.77
NL00641	.30	15.49	.46	13.69	.45	13.79
NL00722	.47	15.67	.63	13.13	.62	13.24
NL00738	.45	16.85	.59	14.54	.59	14.50
NL00807	.46	16.68	.62	14.01	.62	14.08
NL00818	.48	15.98	.59	14.33	.58	14.39
NL00918	.45	15.98	.55	14.55	.55	14.59
NL00929	.41	14.79	.55	12.95	.55	12.89
NL00934	.35	16.20	.45	14.80	.45	15.00
Overall	.42	16.59	.53	14.94		

As noted earlier, further validation was done by applying both the base model and the full NL model to 34 sites in Belgium, not used in model development. Results are summarised in Table 7.9. As this shows, performance statistics are very similar to those obtained in the Netherlands, with R^2 for the full model ranging from 0.43 to 0.68 (average 0.56) and RMSE from 12.3 to 17.6 $\mu g/m^3$ (mean =15.87). Poorest performance tends to be at the western and eastern extremities of the country as demonstrated in Figure 7.5: at two coastal sites and two sites in the hillier Ardennes area of Belgium Significantly, model performance is strongly correlated with distance from the meteorological site, as shown in Figure 7.6. . This indicates that, notwithstanding the results of the sensitivity analysis in the Netherlands (where distance to the meteorological site was always less than 99 km), the model does benefit from having localised meteorological data.

Table 7.9 Performance of NL full model in 34 Belgium O_3 monitoring sites

	Daily Base model		Daily Full model		
Monitoring site	R^2	RMSE	R^2	RMSE	Distance*
BETB004	0.55	14.61	0.68	12.28	89.56
BETB006	0.48	17.36	0.63	14.75	89.43
BETB011	0.44	17.80	0.61	14.74	90.59
BETM705	0.40	16.76	0.52	15.03	142.39
BETN012	0.35	17.83	0.47	16.15	114.37
BETN016	0.44	19.06	0.62	15.81	40.43
BETN035	0.44	19.71	0.61	16.45	65.99
BETN029	0.30	17.83	0.43	16.48	177.87
BETN040	0.42	18.36	0.58	15.70	101.85
BETN043	0.47	14.75	0.63	12.44	85.16
BETN046	0.43	18.16	0.59	15.35	89.59
BETN051	0.43	17.20	0.59	14.66	110.09
BETN054	0.42	18.64	0.56	16.21	96.06
BETN070	0.37	16.13	0.55	13.69	140.60
BETN066	0.33	19.56	0.49	16.08	128.60
BETN073	0.38	17.72	0.54	15.23	118.77
BETN085	0.35	18.52	0.48	16.58	159.67
BETN093	0.37	18.03	0.51	15.83	145.11
BETN100	0.38	19.00	0.52	16.76	165.69
BETN113	0.41	18.89	0.51	17.24	174.57
BETN121	0.41	18.53	0.53	16.58	189.01
BETN132	0.42	19.29	0.52	17.55	211.67
BETR001	0.48	15.42	0.63	13.05	90.04
BETR012	0.43	18.90	0.60	15.72	94.94
BETR201	0.43	17.24	0.57	14.99	113.85
BETR240	0.42	17.49	0.56	15.25	114.14
BETR502	0.39	17.93	0.54	15.51	131.18
BETR701	0.50	15.89	0.65	13.39	101.01
BETR710	0.47	17.09	0.63	14.28	98.39
BETR740	0.41	15.76	0.56	13.57	90.90
BETR801	0.49	15.71	0.64	13.33	52.79
BETR811	0.46	17.98	0.64	14.68	46.76
BETR831	0.44	15.84	0.59	13.50	47.52
BETWOL1	0.40	16.16	0.57	13.70	86.23
Overall	0.45	18.26	0.59	15.87	

- **Distance between Belgium sites and NL meteorological station.**
- **Correlation between the distance and the R^2 is -0.73 and with RMSE 0.59**

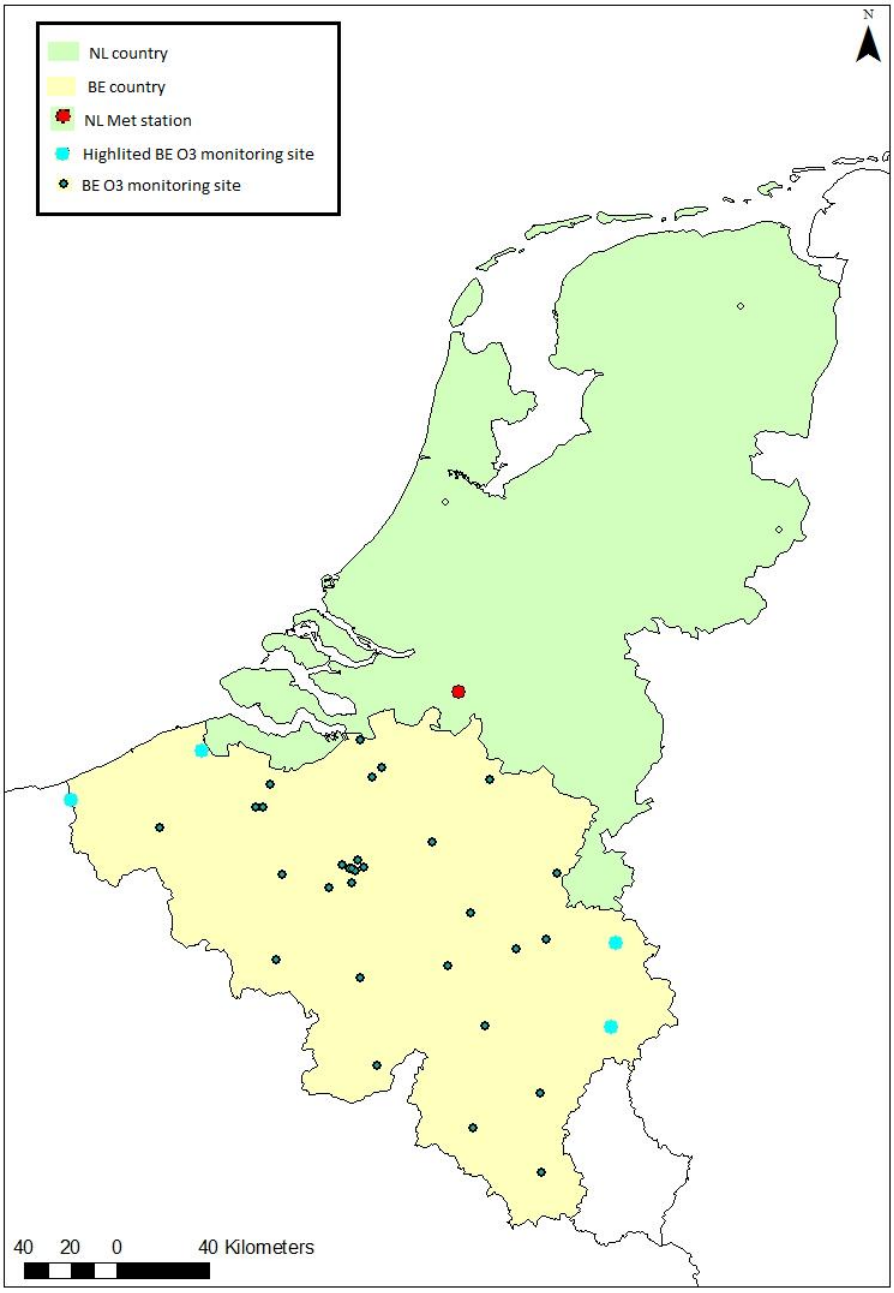


Figure 7.5 Map depicting the location of the Belgium sites and the NL meteorological station

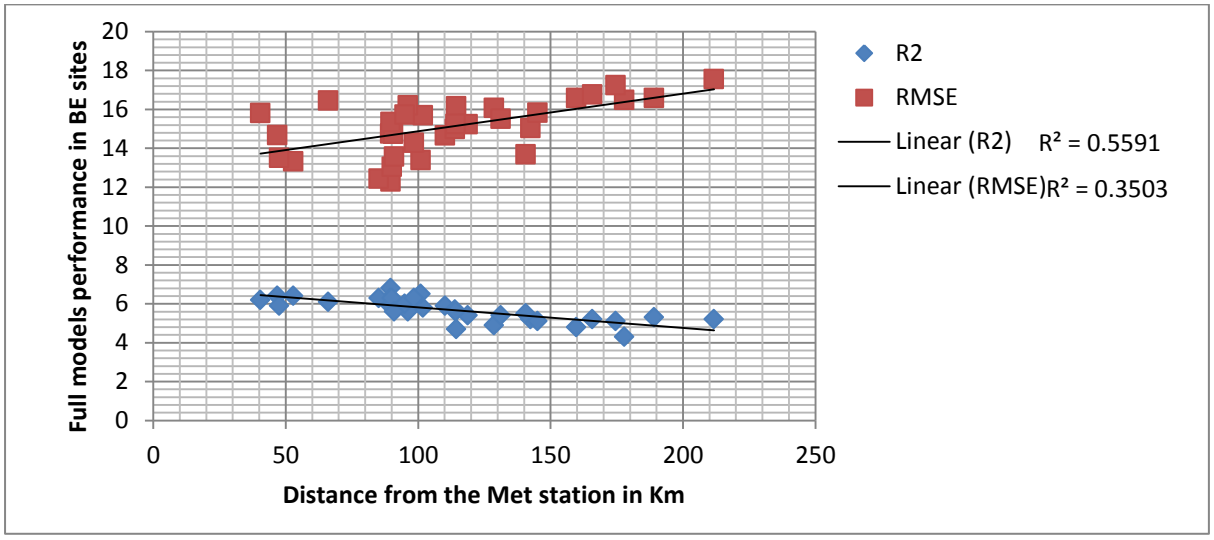


Figure 7.6 Scatterplot between the full model performance measures in Belgium sites and the distance from the meteorological station

The R^2 for the full model is quite strongly and negatively correlated with distance (R^2=0.56), while the RMSE is positively correlated and somewhat more weakly (R^2=0.35). Notably, also, the R^2 for the base models shows a moderate correlation with distance from the meteorological station (R^2=0.37). This may suggest that to some extent there are also other effects from different topography and emission types – Belgium, tends to become hillier and less intensively developed towards the south (i.e. further away from the Netherlands).

In order to explore these results further in NL, the correlation between the performance of the full model (R^2 and RMSE) and the location (x,y co-ordinates) and various environmental attributes of the 1Km area surrounding each of the monitoring sites was analysed. Table 7.10 summarises the results. It is apparent that there is a strong and positive correlation between the percentage of explained variation in the full model and distance to sea (R=0.65); and weaker positive associations with longitude (X-co) (R=36) and altitude (R=0.36). In contrast, the level of explanation falls with latitude (Y-Co) (R=-0.39). As is to be expected, the reverse patterns are seen with RMSE. Given the geography of the country, these associations are all consistent and suggest improving predictions south-eastwards and inland, as shown in Figure 7.7. The reasons for this are not entirely clear. It is possible that the model is not adequately accounting for the effect of O_3 produced over the North Sea, which affects the coastal areas; equally, the lack of factors representing long-range transport of O_3 and its precursors may mean that the model is not adequately reflecting the impact of pollutants carried eastwards over the Netherlands on prevailing winds from the UK or beyond.

Table 7.10 Correlation (R) between performance of the daily full model and different environmental attributes of the monitoring sites

	Daily Full model	
variable	Adj R^2	RMSE
X-Co	0.36*	-0.34
Y-Co	-0.39*	0.32
Urban %	0.22	-0.25
Distance to sea	0.65**	-0.52**
Rural %	-0.16	0.11
Altitude	0.36*	-0.38*

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

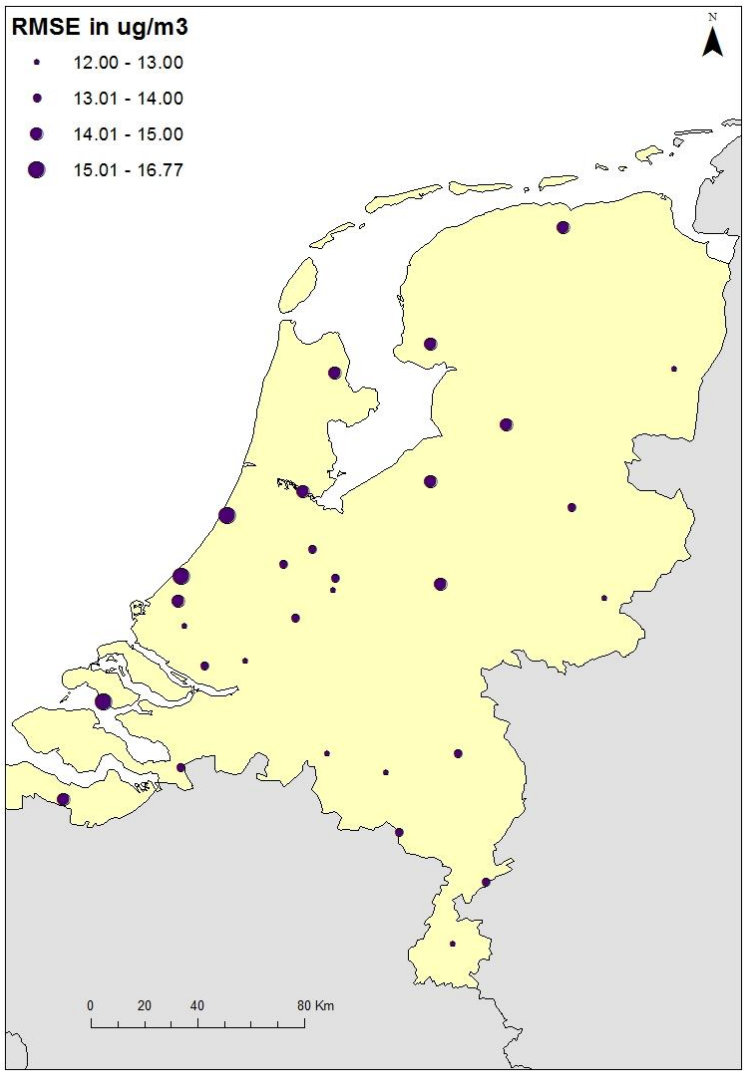


Figure 7.7 Proportional circles of the residual error (RMSE) at each monitoring sites

Table 7.11 shows the effects of aggregating the predicted concentrations to higher temporal scales – i.e. weekly and monthly. Performance of both the base model and full model improves with aggregation, but the difference between the two shrinks and, at the monthly level, their performance is identical, with R^2 = 0.88 and RMSE = 7.62 μg/m^3. This shows that most of the non-systematic variation in the data is short-term and that at the monthly scale systematic variation dominates.

Table 7.11 Comparison between three time scales for the full model in terms of the correlation between observed and predicted concentrations

Model	Aggregation scale	R	Adj. R^2	RMSE
Base model	Daily	0.65	0.42	16.60
	Weekly	0.77	0.60	11.46
	Monthly	0.88	0.77	7.62
Final model	Daily	0.73	0.53	14.94
	Weekly	0.81	0.66	10.54
	Monthly	0.88	0.77	7.62

7.3 Full space-time model in Rome, Italy (case study 2)

Rome is the capital of Italy, located in the central-western portion of the Italian Peninsula, as shown in Figure 7.7. Geographically, Rome covers an area of 1,283km^2 and is inhabited by 2,761,477[27] people according to the latest estimate of Statistics Italy (ISTAT)[28]. Rome is characterized by a Mediterranean climate with a dry summer, with the highest temperature during August reaching 30°C, and a mild, wet winter with rare snowfall.

The GASPII birth cohort consists of 713 participants who were enrolled between June 2003 and October 2004, in two hospitals in the district of a Local Health Unit in the North of Rome, and who were followed up until 2007 (Porta and Fantini, 2007).

[27] http://demo.istat.it
[28] http://www.cbs.nl/en-GB/menu/themas/bevolking/cijfers/extra/bevolkingsteller.html

7.3.1 Methodology

7.3.1.1 Selection of meteorological data

In Rome, the meteorological station in Ciampino municipality (41.8°, 12.55°) was selected as this was located in the vicinity of the five O_3 monitoring sites distributed around the city, and the location of the GASPII cohort participant, as seen in Figure 7.6. Measured meteorological factors included temperature (daily mean °C), wind speed (daily mean $m.s^{-1}$), total cloud cover (oktas), and total precipitation (daily mm) though the last of these had less than 75% data capture so was excluded from the analysis. The descriptive statistics for the meteorological factors available for Rome are summarised in Table 7.12.

Table 7.12 Descriptive statistics for the selected meteorological factors

Meteorological factor	Min	Max	Mean
Wind speed (m/s)	0	10	2.4
Temperature (°C)	-0.4	30.9	15.8
Total cloud cover (oktas[1])	0	8	3.3

[1] In meteorology, an okta is a unit of measurement used to describe cloud cover. This runs from 1 okta (clear sky) to 8 oktas (complete cloud cover)

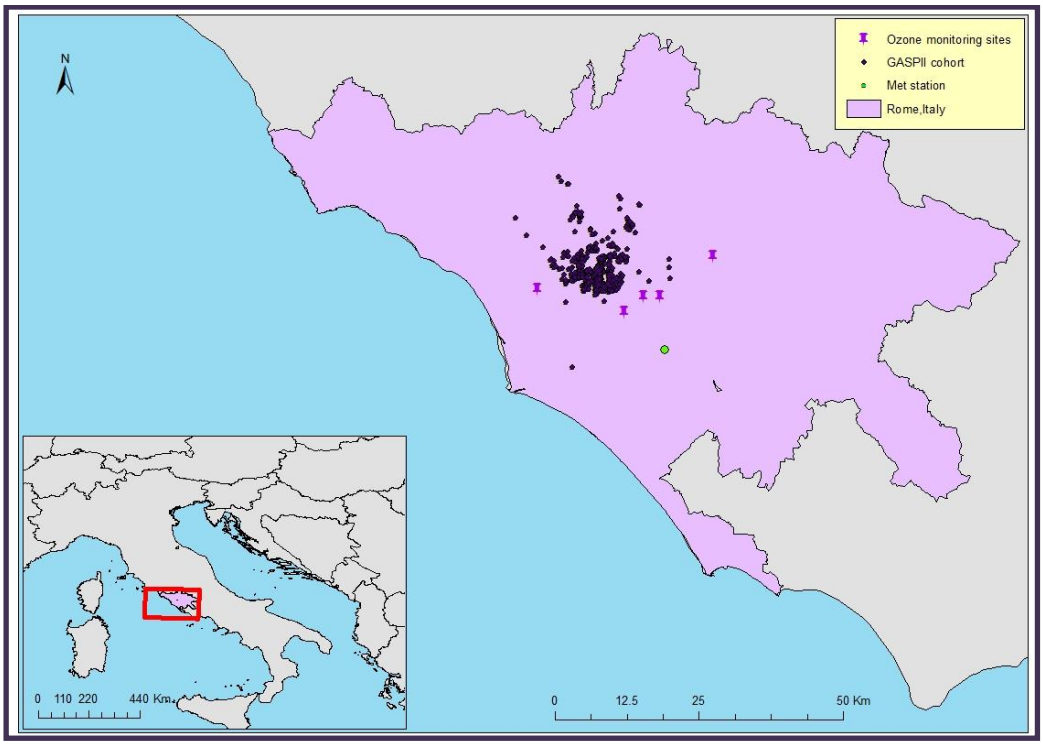


Figure 7.8 Map of Rome showing the locations of the meteorological station, the five O_3 monitoring stations and participants in the GASPII cohort

7.3.1.2 Constructions of full space-time model

To calculate the daily O_3 concentrations for each cohort participant in Rome, the base model (i.e. LUR + daily WTM) was entered into a regression analysis along with the meteorological data, as for the Netherlands. In this case, data for five O_3 monitoring stations were used along with the single, selected meteorological station. Rules for model building and variable selection were as for the Netherlands (case study 1). The model was applied directly to the locations of each cohort participant by using the MLOR equation to calculate the probabilities of site-type membership, and then using these as weighting factors to estimate an appropriate time function model at each location. This was added to the long term mean O_3 concentration produced from the land use regression model. The resulting model prediction and the meteorological factors were weighted and summed using the relevant regression coefficients.

7.3.2 Results and discussion

Results from the multiple regression analysis are shown in Table 7.13. In this case all the meteorological factors added more than 1% to the overall R^2.

Table 7.13 Multiple regression analysis summary of the final model

Model	R	Adjusted R sq.	Change on R sq.	Beta	RMSE	VIF
Base model	0.77	0.59		0.64	15.2	1.69
Wind Speed (WS)	0.79	0.62	0.03	0.23	14.7	1.15
Total cloud cover (TCC)	0.80	0.64	0.02	-0.15	14.3	1.26
Temperature (TEMP)	0.81	0.65	0.01	0.15	14.0	1.77

The regression equation (shown in Equation 7-2) for the daily final model is as follows:

C =-4.07 + (0.79.Base model) + (3.9.WS) + (-1.6.TCC) + (0.48.TEMP) **Equation 7-3**

where C is the predicted concentration, WS is the average daily windspeed, TCC is daily total cloud cover, and TEMP is the average daily temperature.

In terms of the full model, there are positive associations with all meteorological factors except for total cloud cover which had a negative association as expected: a cloudy day will limit the

photochemical reactions occurring due to a reduction in sun radiation. Overall model performance was good, with an R^2 of 0.65 and RMSE of 14 μg/m^3, leaving 35% unexplained variation (Table 7.11).

Table 7.14 shows that adding the meteorological improved the prediction of mean concentrations. However, predictions included a small number of negative values which were substituted with zero, as explained in Subsection 7.2.2.

Model performance was again compared by aggregating the observed and predicted concentrations to weekly and monthly averages. As in the Netherlands, performance improved as the averaging time was extended, with R^2 rising to 0.77 and the RMSE falling to 9.9 μg/m^3 at the monthly level (Table 7.15).

Table 7.14 Descriptive statistics for O_3 concentrations from the daily base model and the full model compared to observed concentrations at the five monitoring sites in Rome

Variable	Minimum	Maximum	Mean	SD
Observed O_3	0	129	40.78	23.65
Base model	3.18	78.77	39.70	18.68
Full model	-3.6	87.88	39.50	19.05
	Daily Base model		**Daily Full model**	
Monitoring sites	**Adj R^2**	**RMSE**	**Adj R^2**	**RMSE**
IT0826A	0.42	13.70	0.66	9.05
IT0828A	0.64	12.13	0.70	11.75
IT0952A	0.46	17.30	0.59	14.30
IT0957A	0.65	13.26	0.75	11.23
IT1174A	0.68	13.32	0.76	11.35

All concentrations in μg/m^3

Table 7.15 Comparison between three time scales for the full model in terms of the correlation between observed and predicted concentrations

Aggregation scale	R	Adj R^2	RMSE
Daily	0.81	0.65	14.00
Weekly	0.85	0.73	11.20
Monthly	0.88	0.77	9.89

The model was applied to the locations of each of the 713 cohort participants, using the probabilities of site-type membership (derived from the MLOR equation) as weights for the different time functions, then adding the resulting daily pollution increments to the long term mean O_3

concentration from the land use regression model. The full space-time model was applied using Equation 7-3 to calculate daily predicted O_3 concentrations for each participant from January 2003 – March 2007 (as required for the cohort). The range of daily concentrations within the cohort was relatively close to the observed daily concentrations at the five monitoring sites, with a mean of 42.8$\mu g/m^3$ (and a range of 0 to 90 $\mu g/m^3$).

7.4 Exposure assessment

As noted above, the three cohorts in the Netherlands and Rome are concerned with birth outcome. The exposures of concern thus relate to the period of pregnancy, and, more specifically, to risks that may develop during specific trimesters during pregnancy. Exposure assessment thus needs to be both spatially and temporally specific, so that it can take account not only of variations in average concentrations between the homes of different participants, but also of the pollution conditions that prevailed during these specific periods at those locations.

In most time series studies, it is assumed that air pollution rises and falls more-or-less uniformly across a city, so that, though the average levels of exposure vary from day to day, the shape of the exposure distribution remains much the same. If this were true, then detailed modelling of the sort done here would not be required, for exposures on any day could be estimated directly from the mapped average concentration (e.g. the LUR map) and the daily concentration averaged across the routine monitoring sites. To explore whether this was the case, or whether the exposure distributions vary over time and space, a sample of 200 participants was randomly chosen from the PIAMA cohort (described in section 7.2). These 200 participants are scattered across NL but show some degree of concentration in the north, west and central areas of the country (reflecting the distribution seen in the original cohort). For each participant a random set of daily, weekly and monthly averages was then selected. This sampling design thus provides a temporally and spatially unbiased cross-section of the data.

Exposure estimates were then made both from the full model and from the nearest monitoring site. Distance between the monitoring sites and participants range from 0.25 to 31.67 km with a mean of 8.88 km. Computed exposure estimates from the two approaches (full model and nearest site) are presented below in different ways for the different time scales:

1- as histograms to illustrate the distribution of the exposure estimates (mean, SD, and the range) and to show the number of participants in each exposure category (Figure 7.9).

2- as scatterplots, showing also the 1:1 line, to depict the correlation between the two measurements and show if one approach over- (or under-) estimates compared to the other (Figure 7.10).

As these results show, the two approaches give different exposure estimates for each averaging period. At the daily level, the means are approximately the same, as is to be expected because the model is calibrated to the measured mean (ca. 49 ug/m^3) from the 30 monitoring station across the Netherlands. The SD and the range, however, differ (Table 7.16) with results from the full model showing a more restricted range, indicating that fewer people get assigned extremely high or low exposures. As can be seen from the histogram (Figure 7.9) and the skewness scores (Table 7.16), the estimates from the full model are also somewhat less skewed than those from the nearest site.

The correlation between the two daily estimates is also weak (R^2=0.12), showing that they give different exposure estimates to the sample of participants, and there is a marked tendency for the full model to over-estimate compared to the nearest site at low concentrations and to show relative under-estimation at higher concentrations . These differences arise because, using the nearest site approach, concentrations are extrapolated over relatively large areas and all individuals within that area; the full model, on the other hand, attempts to show the spatial variation within these areas. In general, this acts to redistribute exposure scores within the limits of the monitored concentrations.

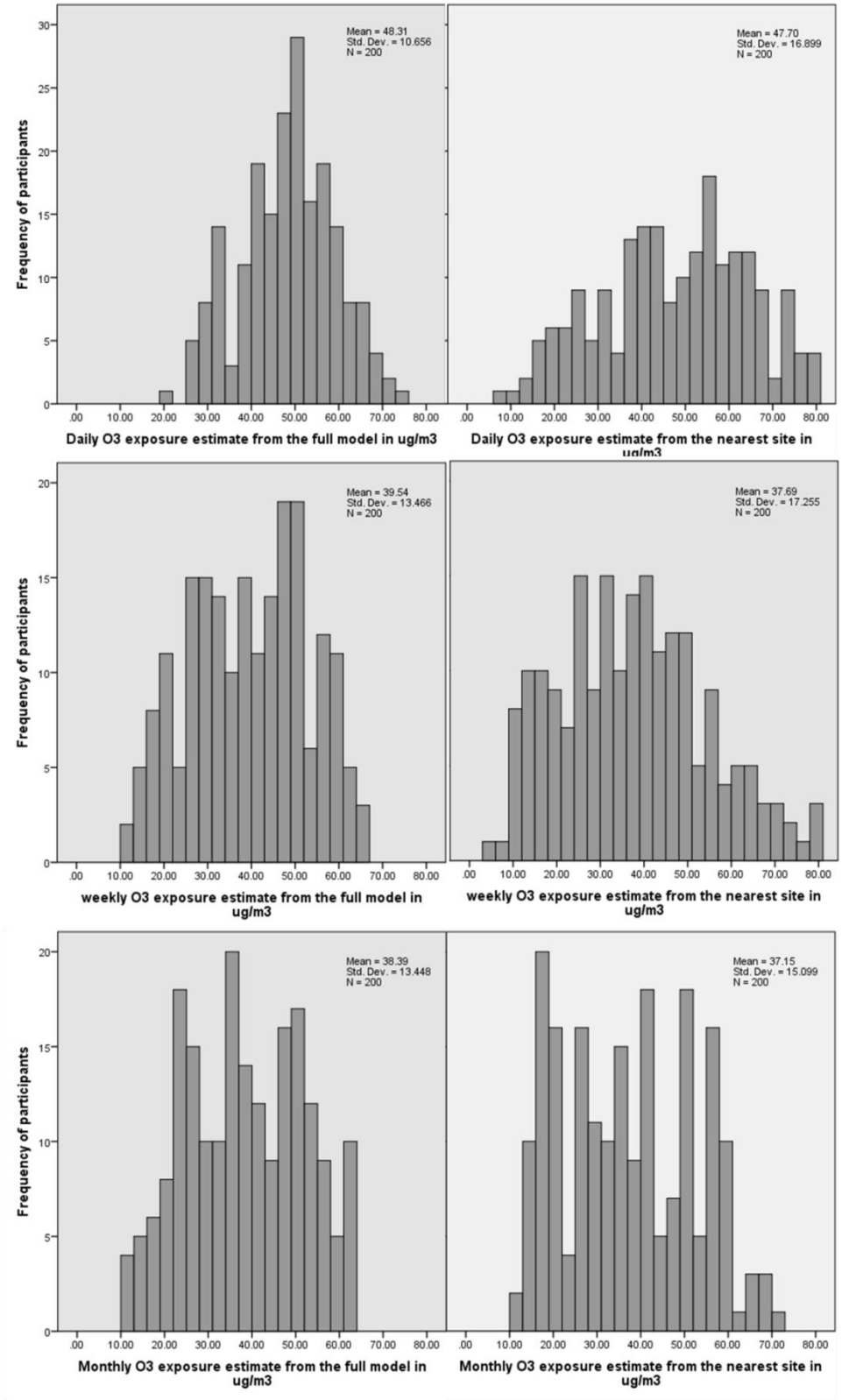


Figure 7.9 Histograms for the exposure estimate distributions for the 200 participants in NL, for different averaging times (daily, weekly and monthly) for the full period of the study using two approaches (model prediction from the full model and nearest monitoring site)

Table 7.16 Summary statistics for each averaging period and approach

Temporal scale	Exposure estimation	Min	Max	SD	Mean	Skewness
Daily	Full model	19.96	73.85	10.65	48.31	-0.14
	Nearest Site	6.06	80.93	16.89	47.69	-0.19
Weekly	Full model	10.22	65.11	13.46	39.54	-0.11
	Nearest Site	4.90	79.86	17.25	37.68	0.34
Monthly	Full model	10.9	63.70	13.44	38.38	0.00
	Nearest Site	12.29	72.63	15.09	37.14	0.20

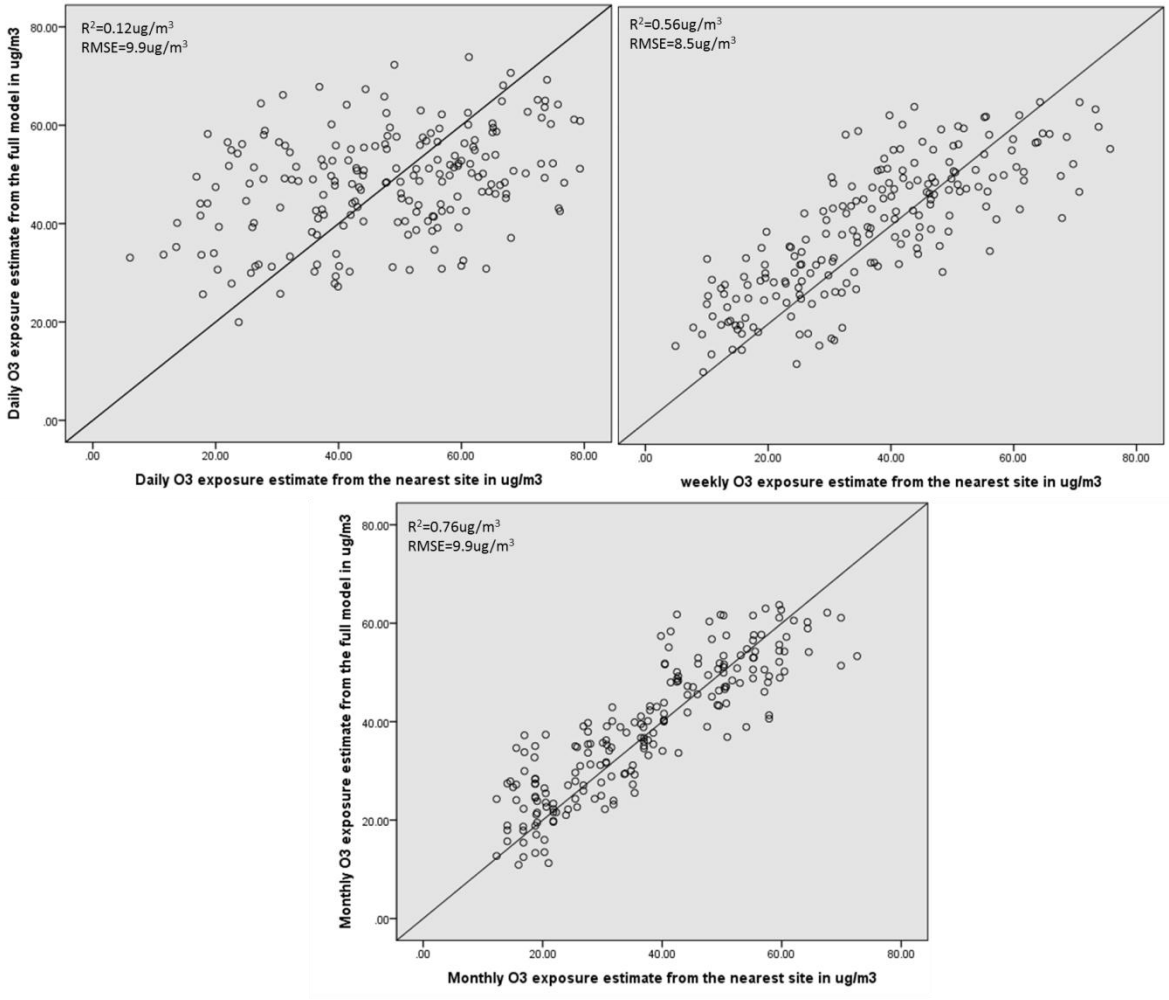


Figure 7.10 Scatterplots (with 1:1 line) of the exposure estimaes from the two approaches (full model and nearest site)

At the longer averaging times (week and month) the exposure estimates from the two methods are much closer. The correlation between the two methods increases to R^2=0.56 at the weekly level, and R^2=0.76 at the monthly level. This is to be expected, for as has been noted, much of the variation in the date occurs over short (daily or hourly) time scales: variance components analysis of the hourly data in the Netherlands, for example, shows that 6% of variation is spatial and 31% is temporal. As the averaging period increase, therefore, much of the variation (from day to day) is removed, and the correlation between the two approaches improves. While the means remain close (mean=39.5 and 37.7 ug/m3 for the weekly estimates and 38.4 and 37.1 ug/m^3 for the monthly for the nearest site and full model respectively), both the standard deviation and skewness differ. For both averaging periods, the nearest site approach has a higher standard deviation and is more positively skewed.

The patterns seen in these data are not necessarily true for other areas, for O_3 concentrations are strongly affected by the geographic patterns of emissions and topography, and the temporal patterns of meteorology. These factors also affect how the different approaches work: for example, in a flat, maritime environment such as the Netherlands, spatial variability in ozone is likely to be reduced compared to that in a topographically or climatically varied environment. Here, therefore, the nearest site is likely to give valid estimates of the true concentration over somewhat larger distances.

To explore this issue further, estimates of exposures for the 713 participants in the Rome cohort were also made using both approaches for different time periods (days, weeks and months). Figure 7.11 shows exposure distributions from the full model for three consecutive days in January 2003, for three consecutive weeks in the same month, and for three consecutive months in 2003. Figure 7.12 shows estimates for one month (March) in three consecutive years. Figure 7.13 shows estimates for the same days, weeks and months in 2003, based on the nearest site); numbers of participants and mean distance to the monitoring site are given in Table 7.17.

Results from both methods show that exposures differ quite markedly over these timescales, not only in terms of the absolute level, but also their distribution. In terms of the day-to-day variations, this is especially apparent on the third day and the third week in January, when the shape of histogram changes from that of the previous two days. Likewise, the week-to-week exposure distributions vary considerably over the first three weeks of that month.

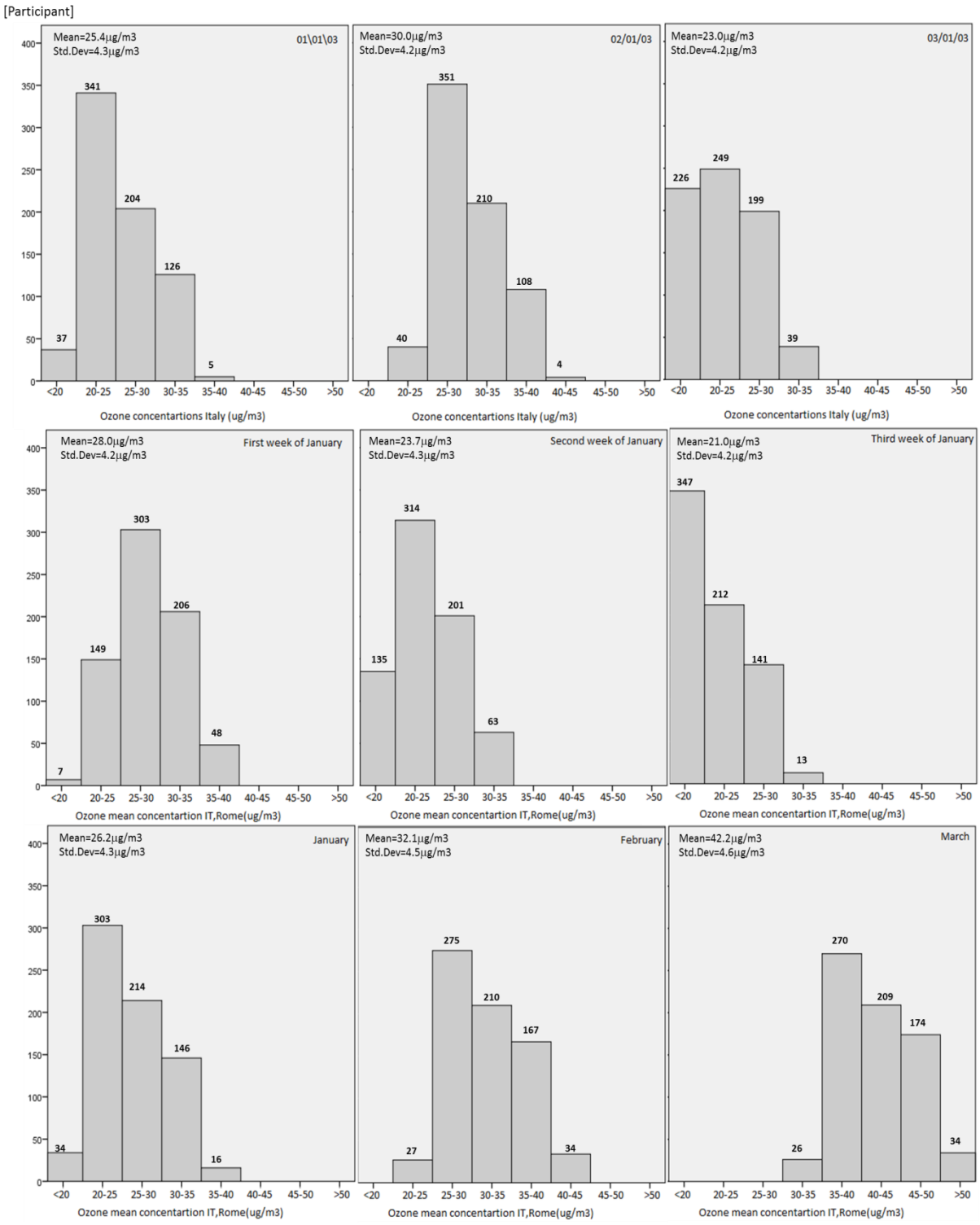


Figure 7.11 Exposure distributions across the 713 cohort participants are estimated by the full model, in Rome, in 2003

At the monthly level, also, both the average concentration and the distribution change between January and March. In the case of the year-to-year differences (full model only), the situation is slightly different. Reflecting differences in weather conditions in March between these years, the mean exposure rises over these three years. The shapes of the histograms, however, remain broadly the same. The consequence is that the number of participants exposed to concentrations above 50 $\mu g/m^3$ rises quite sharply – from 8 in 2004 to 34 in 2005 and 84 in 2006.

Together, these results suggest that the timing of critical exposures (in this case during pregnancy), as well as the location of residence, may have an important influence on exposures, and thus on health outcome. They thus highlight the importance of considering both temporal and spatial variations in concentration when studying health effects for which the critical period of exposure is short.

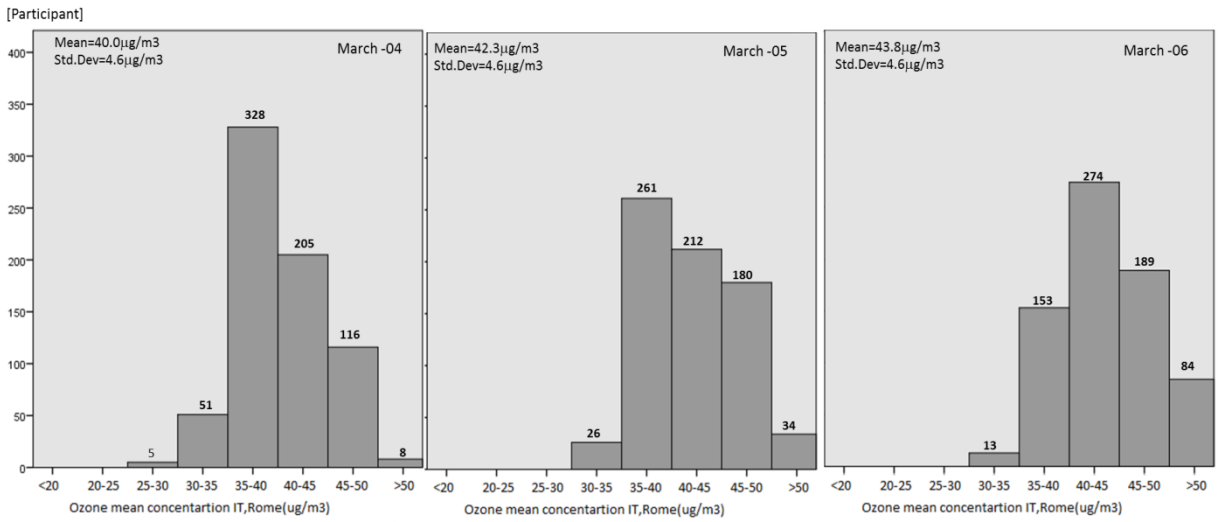


Figure 7.12 Exposure distributions for the 713 cohort participants in Rome in March 2004, 2005, and 2006

Table 7.17 Number of participant assigning to nearest sites and distance measures

Monitoring sites	No. of participants	Distance average (m)
IT0826A	130	8277
IT0828A	198	6497
IT0952A	382	6979
IT0957A	1	2964
IT1174A	2	4451

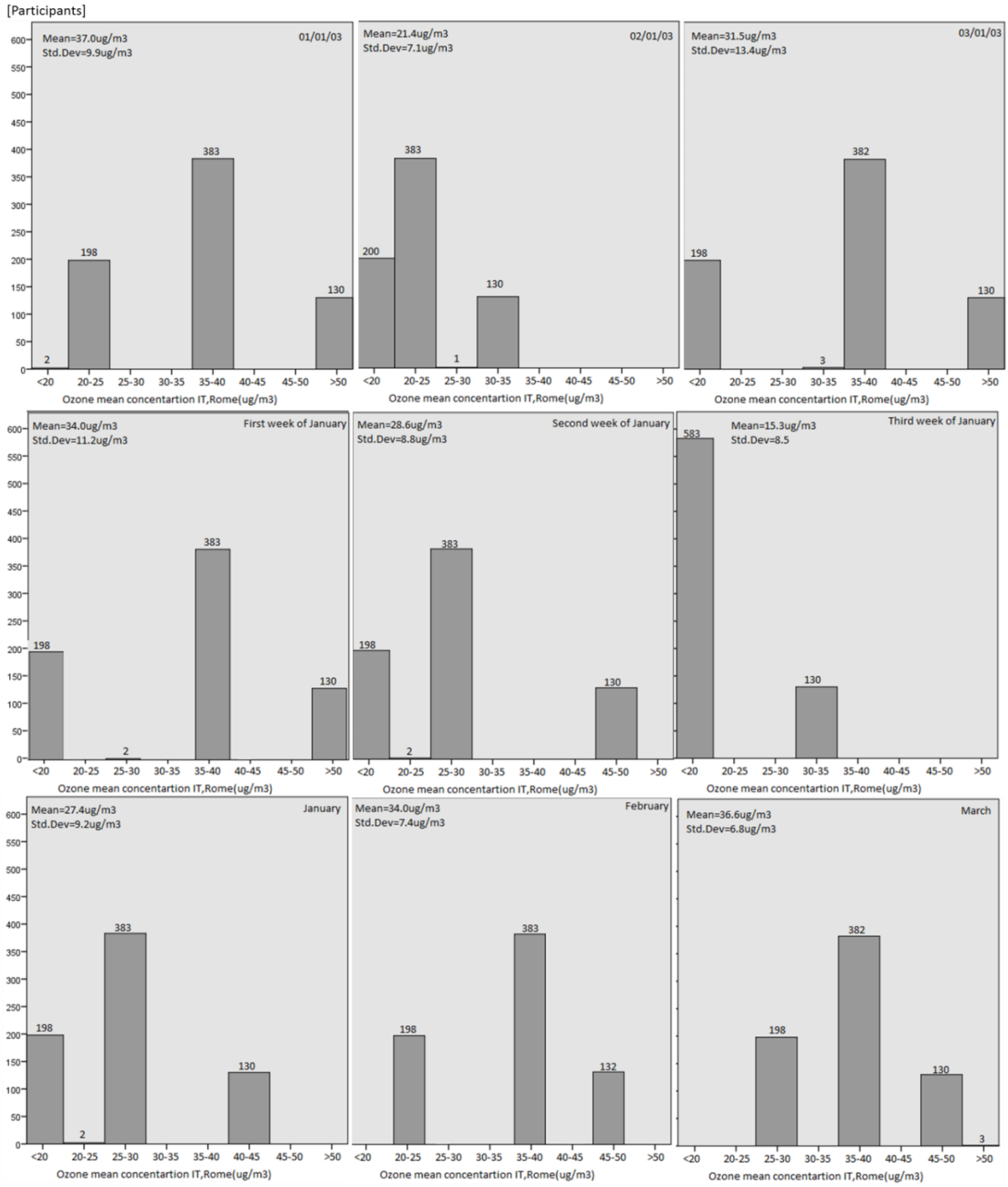


Figure 7.13 Exposure distributions across the 713 cohort participants in Rome, in 2003 by assigning participants to the Five nearest monitoring stations

It is also apparent, however, that the exposure distributions from the two approaches differ considerably. In particular, the estimates from the nearest site are much more fragmented (with gaps between the different exposure scores), simply because the concentration measured at any site, for any specific time interval) is assigned unchanged to all the surrounding participants. This is also reflected in the standard deviations, which are very large for the nearest site compared to the full model. By the same token, the exposure distributions derived from the nearest site are much more variable over time.

7.5 Summary

This chapter has described how, for two different study areas, the spatial and temporal models were combined - first in a base model, then in a full model which attempted to account for non-systematic variation in O_3 concentrations. This non-systematic component of the temporal variation may occur for a number of reasons. These could include different temporal patterns of emissions, especially in urban areas, variable effects of transport of O_3 or its precursors from neighbouring areas, differences in dispersion efficiency and episodic variations in weather conditions. In order to capture this non-systematic variation, daily data for four meteorological factors (temperature, windspeed, sun duration or cloud cover, and total precipitation), were offered into the base space-time model in both the Netherlands and Rome.

While the daily base model explained on average 44% of the variation across the 35 monitoring sites in these two study areas (with RMSE = 16 $\mu g/m^3$), the full model which included these meteorological variables explained 57% of the variation on average (with a range from 45-76%), and with an average RMSE of 14 $\mu g/m^3$ (ca. 28% of the mean concentration). Generally, this shows a good fit of the full model to the observed data, especially in Rome city, and suggests that the results can justifiably be used as a basis for exposure estimation on a daily basis.

Aggregating the full model to weekly and monthly levels, significantly increased the proportion of explained variation (to 77% at monthly level in both areas) and reduced the error in the predictions (to <8 ug/m^3 in NL and <10 ug/m^3 in Rome). The results reflect the fact that weekly, and in particular, monthly O_3 concentrations are the result primarily of seasonal variations in temperature and solar radiation, which are both broader and more systematic in their scale of effect.

Post hoc analyses were also carried out to compare results from the full model with those from the traditional approach of assigning participants to the nearest monitoring site. These showed that,

while the two approaches give similar estimates of mean exposure (as is to be expected because the model is calibrated to the measured mean), the shape of the exposure distribution tends to vary and the individual exposure estimates vary, especially for short averaging periods.

It is not possible to determine which method is more accurate from these data. Nevertheless, for several reasons it may be inferred that the nearest site approach is liable to give less reliable estimates. Firstly, this is because the approach makes no allowance for the local variation in concentrations that is known to occur as a result of the effects of local emission sources, topography and meteorology. Using data on these factors, which are clearly correlated with O_3 concentrations, is likely to improve exposure estimates. Secondly, the exposure distributions generated by the nearest site approach is shown to be much more disjunct: individuals are thus 'forced' into discrete exposure classes, whereas, in reality, exposures are likely to vary much more smoothly. Thirdly, exposures estimated from the nearest site are shown to be much more variable over time, simply because exposures for large numbers of individuals change with the (often discrete) changes at the monitoring site. Timing of measurements (or of the predefined exposure window) may thus have a large effect of exposure estimates, and on any observed association with health outcome. The question of which approach is more reliable is considered further in the next chapter.

8 Discussion

8.1 Modelling principles

Exposure to tropospheric O_3 has been identified as a critical public health concern in Europe, and a number of policies exist, aimed at limiting the production of O_3, largely by controlling the concentrations of O_3 precursors. Recent EU guidance has also been established to help protect human health, as mentioned in Chapter 2 of this thesis.

Despite these measures, O_3 concentrations still exceed the threshold of 120 $\mu g/m^3$ which is considered to represent serious risks to health, in many locations and on many occasions through the year. Meanwhile, new epidemiological studies have indicated that significant health impacts may arise at concentrations less than the concentrations specified by the current standards.

Policy makers establish O_3 guidelines based on epidemiological studies, which in turn rely on methods of exposure assessment to define the concentrations at which detectable health effects occur. Many factors confound and complicate the relationships between health and exposures to O_3, including the duration of exposure, the age of participants, and the medical history of the subjects. Individual-level studies are essential to unravel these complexities. At the individual level, however, effects may be subtle and risks may be small.

Large, longitudinal studies are therefore needed to assess these risks with any degree of reliability. Cohort studies, involving participants representing different age groups, locations (i.e. different places, cities, and countries) and different exposure histories, who can be followed up for a relatively long period of time, offer the most effective study design. The challenge for exposure assessment is to provide the sorts of data that these studies imply.

Methods for estimating such exposures have generally been lacking. Traditional, ground-based monitoring is expensive, and routine monitoring networks have not been designed to represent exposures with this degree of precision. Satellite-based measurements are beginning to provide the capability to map tropospheric O_3 concentrations over time, but their spatial resolution (in both the horizontal and vertical dimensions) is currently inadequate for the purposes of epidemiological analysis. Methods to model O_3 concentrations have been developed, but these are generally too demanding in terms of data and processing capacity to be usable for exposure assessment. While a few attempts have been made to model spatial patterns of O_3 over the whole of Europe (Beelen

et.al, 2009), no-one, to the knowledge of the researcher, has yet developed a space-time model at this fine scale (100m*100m) of analysis that meets the requirements for epidemiological studies. Those that do exist (e.g. Bruno et.al, 2009, Gariazzo et.al, 2007, Moolgavkar, 2000) tend to be specific to individual cities, regions or countries.

The question of this thesis was therefore whether it is possible to produce a generic and usable GIS-based methodology for modelling spatial and temporal variations in O_3 concentrations over a large (pan-European) study area, at a spatial and temporal resolution appropriate for epidemiological studies and health risk assessment in support of policy. The approach taken was to develop a three-part model, comprising:

1. A spatial model, based on land use regression, that can predict the long-term average O_3 concentration at a resolution of ca. 100 metres, using readily-available geographic data, (Chapter 4).
2. A set of time-functions, for different site types, that could describe systematic temporal variations in O_3 concentrations over averaging periods ranging from hours to several months (i.e. seasons), (Chapter 6).
3. A regression-based model to take account of non-systematic temporally correlated variations due to changing weather conditions, using readily available meteorological data, (Chapter 7).

The first two of these models were developed and calibrated using data from 1211 O_3 monitoring sites across Europe (stratified into 80% training and 20% validation subsets). The third component of the modelling was undertaken in two smaller study areas (the Netherlands and the city of Rome), both of which contained cohorts feeding into the EU-funded ESCAPE project. This analysis was done both to demonstrate how adding meteorological information might enhance the model, and to illustrate the potential for application of the full model as part of an epidemiological/health impact study. This, therefore, provided an opportunity to evaluate the overall approach as a basis for exposure assessment in cohort studies.

All three models were built on a common underpinning theory of spatial and temporal variations in pollutant concentrations: namely that variation in both space and time comprises three main components:

- A systematic, and repeated pattern of variation (sometimes known as trend or drift);

- A non-systematic variation, associated with measurable, exogenous factors that vary over time or space – referred to here as the random, spatially (or temporally) correlated variation;
- A non-systematic, truly random variation that is not correlated with measurable exogenous factors, and thus represents noise.

All three elements of the model were designed explicitly to reflect the factors and processes known to influence O_3 concentrations in the atmosphere. The model thus broadly reflects the main components found in standard dispersion models. The goal, however, was to do so in a way which allowed the models to be more easily used with readily available data and with limited computer power.

8.2 Categorizing O_3 monitoring sites

Different temporal models were developed for each of categorise of site type. Classification, however of sites based on temporal pattern is not a straightforward process and needs to be informed by a thorough understanding of the factors that contribute to O_3 variation over time. Using contextual environmental characteristics to define site types has many advantages due to the stability of these factors over time. This is therefore the approach usually taken to characterise site types, as in the AIRBASE classification, for example. Clear distinction is generally possible between major categories of site types defined in this way: rural and urban types can, in particular, be readily distinguished on the basis of land cover data (e.g. Beelen et al. 2009).

For primary pollutants, such as NOx or traffic-related particulates, this approach works successfully, and the different site types generally represent situations that differ in terms not only of average pollutant concentrations but also their temporal signatures. This is because the factors used to define site types are directly involved in controlling emissions. For a secondary pollutant such as O_3, the approach is likely to be less effective, for the association of the stable, environmental characteristics with pollutant concentration is far less direct, and more complex. In the case of O_3, photochemical reactions play the determining role, and these depend not so much on static characteristics of the environment, such as land cover, as time-varying meteorological factors. Data on these are rarely if ever available for a dense network of locations, suitable for defining site types,

but instead come from a sparse network of measurement stations, or as regional scale variables derived wherefrom.

The approach taken here was thus to characterise site types on the basis of the temporal signatures of O_3 concentrations, and then to seek ways of predicting site type membership using exogenous, environmental characteristics. Initially, analysis was done to explore the patterns of temporal variability in the data, and to select a series of temporal indicators that could characterise this variability. These temporal indicators were grouped into the four components, using PCA, and these then used to define site types via HCA. Thirteen site types were thereby identified, varying in terms of the patterns of pollution over three main timescales: weekday/weekend; summer/winter; and night/afternoon.

The thirteen site types derived from the HCA cannot be seen as definitive and discrete entities. Instead, it is apparent that the different types overlap and merge into each other, and their classification involves considerable uncertainty. Nevertheless, classifying site types on the basis of the temporal variability is likely to be preferable to basing classification solely on environmental factors as Flemming et al. (2005), Joly and Peuch (2012) and Snel (2004) have previously argued. Site type membership, however, must be recognised as probabilistic, and is highly dependent on the indicators selected for analysis. For these reasons, great care is needed both in applying the classification in the study area, and even more so in extrapolating it to other areas or time periods. Testing and validation of the classification is essential.

In most situations – as here – analysis does not stop with the creation of a classification. The further need usually exists to apply this classification to other, unmonitored sites, in order to model or map the behaviour of the pollutant of interest. It is therefore necessary to find some means of predicting site type membership at these unmonitored sites. The next step in this approach is therefore to establish relationships between the site types and other, exogenous environmental variables that can be measured at the unsampled locations. This was done here using MLOR analysis with a series of variables relating to land cover, roads, meteorology and topography as predictors. The resulting functions have two purposes. They reveal the environmental variables that are associated with (and directly or indirectly act to influence) temporal variations in O_3 concentrations; and they provide the basis on which to predict site type membership elsewhere.

The functions established were dominated by four main sets of environmental variables: local roads and residential land, which reflect the distribution and intensity of the main precursor pollutants; temperature, which determines the efficiency of photochemical reactions; windspeed and topex

which act to determine air exchange (and thus the ingress and egress of both precursors and O_3) between any location and its surrounding area; distance from seas, which reflects the influence of marine sources of O_3; and altitude which acts as a complex co-determinant of meteorological conditions (e.g. sunshine, temperature, wind speed and turbulence) and tropospheric exchange with the stratosphere, as well as a proxy for remoteness from emission sources.

Again, however, the uncertainties in this analysis need to be recognised. In this case, only 52% of the sites were classified by MLOR (on the basis of environmental variables) to the same category as their initial classification using HCA (based on the pollutant signatures). As Figure 8.1 thus demonstrates, the different site types do not exist as distinct entities in multivariate space, but tend to cluster; some degree of confusion in the classification is therefore inevitable.

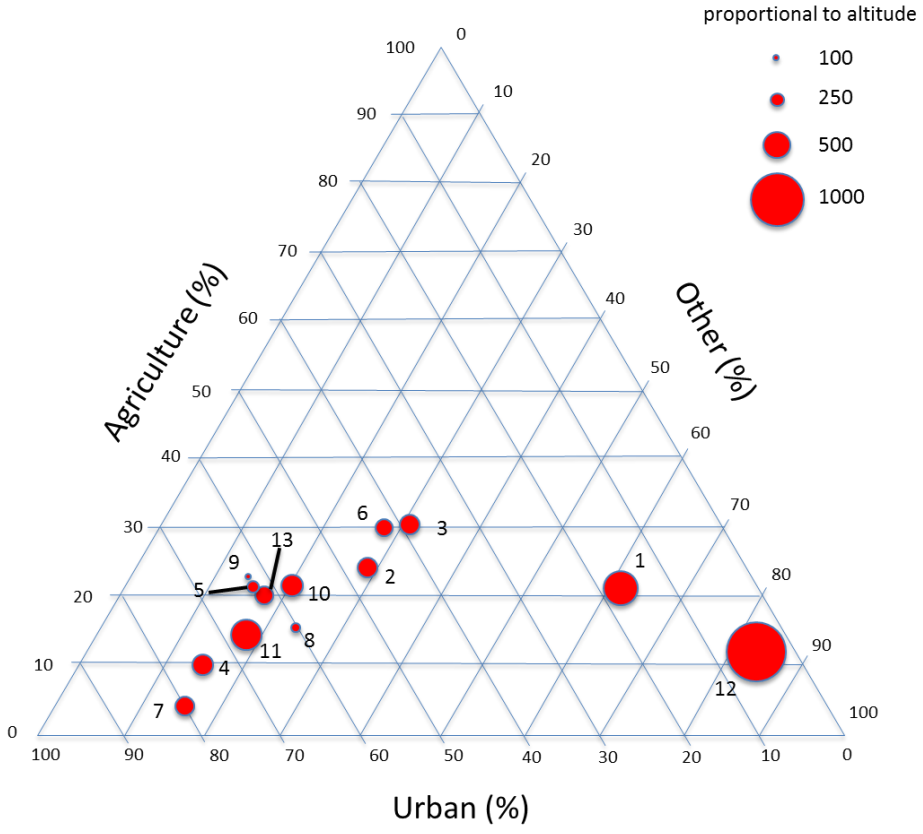


Figure 8.1 Triangular plot for the percentage of urban, agriculture, and other land, and altitude (metres), for the thirteen site types

This problem is not unique to the analysis conducted here. The same problem was noted by Joly and Peuch, (2012), who likewise first categorised sites on the basis of temporal indicators, and then derived a discriminant function to distinguish between rural and other (urban, suburban, and

traffics) sites, as defined in AIRBASE. The result showed only a weak match, with both the rural and urban AIRBASE sites broadly dispersed across their ten site types.

This lack of a clear relationship between the temporal characteristics of air pollution, and the fixed, spatial characteristics of the environment is not surprising, but should not be ignored. On the one hand it is not feasible to ignore the temporal variations in air pollution, if interest is in exposure assessment. On the other hand, it should not be assumed that these variations are either consistent over space, or are readily predictable on the basis of spatial features. Instead, the temporal behaviour of air pollutants (and especially secondary pollutants such as O_3) needs to be recognised as a geographically fuzzy process, which needs to be dealt with using appropriate fuzzy analytical methods (i.e. probabilistic approach).

8.3 LUR model for secondary pollutant

The spatial modal was built across the Western Europe based on the fact that the land cover data will explain most of the spatial variation between the different site types. For example, the residential land cover classes will reflect how the O_3 concentrations will be in the urban area whereas the forest and agriculture will reflect the O_3 concentration in rural area.

As already noted, most previous LUR models have focused on modelling primary (or near-primary) pollutants, and these models are driven mainly by the distribution of the emission sources of the pollutant of interest. In modelling O_3, LUR is being used to predict the distribution of a secondary pollutant which is the product of a more complex set of processes, operating over timescales of several hours. In this case, therefore, the emission sources of interest are one-stage removed: they relate to the precursor pollutants that control production and loss of O_3. Using LUR in this way is likely to be applicable so long as the control provided by these pollutants is reasonably direct and explicit (i.e. the chemistry does not depend on a large number of other contingent factors) and data on their source distribution and intensity are available.

The model developed here used several different variables to represent the main processes and circumstantial factors expected to control O_3 generation and loss. The relationship between these variables and the processes they represent, and their collective influence in the model, can be summarised as in Figure 8.2.

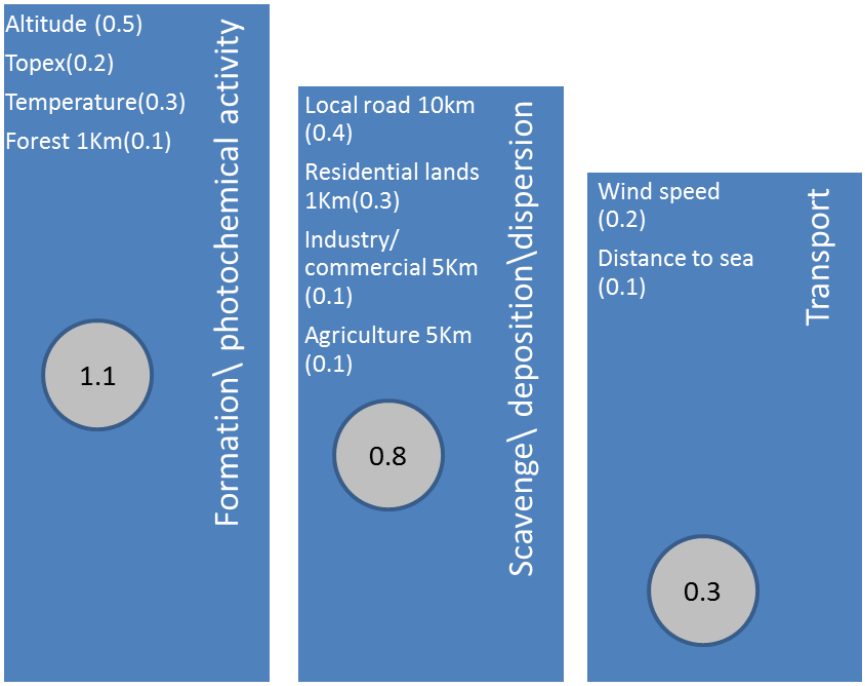


Figure 8.2 The importance the main processes and factors controlling O_3 generation in the LUR model, as shown by the sum of the standardised Beta coefficients for related variables (in the circle)

Based on Figure 8.2 it is clear that the model represents all three of the main processes determining O_3 concentrations – formation, destruction and transport. As is to be expected (because the effects are local and relatively direct), the model seems to load most highly on processes relating to O_3 formation: i.e. local topography and temperature (all of which are primarily indicators of for photochemical activity), and forest area (which is a source of VOCs). The fact that altitude and topex come into the model, rather than solar radiation, is likely to be because altitude is better measured, and spatially more precise, than the measures of solar radiation, which are based on broader-scale data. Models might thus be improved by having higher resolution metrological data available.

Destruction and loss are also implied by variables having a negative effect in the models, most of which relate to sources of emissions of NO (a major scavenger). Agricultural land is the exception, and can be tentatively interpreted as an indicator of O_3 removal by dry deposition. Inward transport of O_3 is indicated by two variables – windspeed and distance to sea – but the effects of both are relatively weak. This is partly, perhaps, because these processes operate over a broader scale and

are often transient and circumstantial (e.g. dependent on wind direction). However, both variables are also relatively poorly measured compared with the traffic and land cover variables. For instance, the original data of the meteorological factors were on a 40Km grid, having been previously derived through a modelling and estimation procedure from ground-based and satellite observations. For this study they were downscaled to a 100m grid using a simple interpolation procedure (IDW). This inevitably involves considerable approximation. At best, therefore, the meteorological data reflect regional effects rather than local processes. None of the variability due to local topography, urban area or individual buildings is therefore allowed for.

The role of wind speed deserves special comment. In general, there might be an expectation that windspeed should be negatively associated with O_3 concentrations because higher wind speeds encourage dispersion and mixing of pollutants (Hubbard and Cobourn, 1998, Bloomfield et al., 1996). This expectation certainly tends to hold true for primary pollutants or secondary pollutants produced from direct simple reactions, but it may be less valid for secondary pollutants produced by a number of complex reactions such as O_3. Some other studies have thus found positive associations with wind speed (Elampari and Chithambarathanu, 2011, Shan et al., 2009, Tarasova and Karpetchko, 2003, Davies et al., 1992). Shan et al. (2009) explained the positive association in two ways. One is that wind transports O_3 and precursors from the direction of highly polluted areas of the city, so higher wind speeds represent higher rates of input. The second is that higher wind speed indicates the long-range transport of air pollutants. Both of these, of course, assume that the monitoring site in question lies downwind of the sources – a condition that is unlikely to hold true across all sites. Davies et al. (1992) also showed that the effect of windspeed varies across the year. In winter, the relationship with O_3 concentration became strongly positive, especially in areas more influenced by the prevailing wind. In spring and autumn it exhibited either a negative or a positive relationship depending on geographical location. They attribute the positive relationship to the importance of vertical exchange from the free troposphere to the surface in non-summer months – a process which is encouraged under turbulent conditions and high windspeed.

Again, therefore, models might be improved by having variables which better represent transport of O_3 – for example, on the prevailing wind direction relative to major O_3 sources, or on O_3 concentrations in upwind directions. Likewise, a potentially important variable that is missing is one that represents the potential for the replenishment of surface O_3 from higher layers in the atmosphere (e.g. the stability class or an indication of the vertical profile of O_3).

8.4 Comparisons with other studies

8.4.1 The LUR model

Most previous studies using LUR to predict air pollutants have been done at the local scale, within cities, or within single countries. These have generally shown that LUR can effectively predict intra-urban variations in air pollution as described in section 2.2.2.4. Likewise, most previous applications of LUR have been to primary pollutants or pollutants that form quickly and close to the emission sources of their precursors, such as particles and NO_2. The results presented here demonstrate how LUR can be applied at high resolution even across an entire continent. The R^2 value was similar to those found in local scale studies (Hoek et.al, 2008), and the few continental scale studies that exist (Beelen et al., 2009, Vienneau et al., 2009, Nikiforov et al., 1998).

Opportunities to directly compare these results with those from other O_3 studies are limited because only a few previous studies have modelled O_3 at a broad (e.g. continental) scale. Nikiforov et al. (1998) modelled long-term O_3 concentrations (10 year mean) using the average of O_3 data measured using five different metrics[29] from 1112 monitoring sites over the whole of the U.S.A. They also used five different interpolation methods: a) simple average from nearby sites; b) inverse distance weighted interpolation; c) inverse distance squared weighted interpolation; d) regression analysis; and e) ordinary kriging. Although not directly comparable to LUR, the best results were obtained using a regression-based interpolation on the three nearest sites (R^2 ~ 0.7and SEE=15.8 µg/m^3) based on 1- hour maximum concentrations. Note that distances between monitoring station were not mentioned in the paper.

Beelen et al. (2009) modelled and mapped the annual average O_3 concentrations at a 1km level for Europe using data from 724 monitoring sites, representing measured background concentration in rural and urban environments, by linking models developed at three different scales (i.e. global, rural and urban). Three different methods were tested: a) ordinary kriging; b) universal co-kriging and c) land use regression. They found that universal co-kriging (using covariate data based on the LUR model) performed better than ordinary kriging and LUR alone. As per this study, validation was done using a separate subset of the O_3 data reserved for validation. The final validation results for universal co-kriging estimates gave R^2 = 0.70 and RMSE = 7.7 µg/m^3 while for LUR the results were R^2 = 0.53, 0.62, and 0.06 for global, rural, and urban scales respectively. Based on universal kriging, the

[29] Average concentrations of 1 hour maximum,8-hour maximum, daily concentration between 10am-10pm,daily average between 10am-6pm, and summer hourly average ≥ 120 µg/m^3

range of predicted concentrations in the final map was from 12μg/m^3 to 153μg/m^3, with a mean of 60μg/m^3 while the observed concentration in the training dataset was 19-112 μg/m^3, with 49 μg/m^3 mean. The kriging model therefore, overestimated the mean. Results from the present study are broadly similar (R^2 =0.67 and RMSE = 7.5 μg/m^3), but this study has included all site types rather than just background ones, as in Beelen et al. (2009). The range in the modelled data of this thesis' study was from 7.9μg/m^3 to 182.6μg/m^3, with a mean predicted concentration of 49.6μg/m^3, and the observed concentration was 15-111 μg/m^3, with 49.6 μg/m^3, indicating no bias in predicting the mean.

Eleven significant predictors were found to be important for this global LUR (spatial) model, this number of predictor variables reflect the different process contributes to O_3 concentration (Figure 8.2). As noted, the first seven predictors in terms of importance (Figure 5.2) were altitude, local road density within a 10km window, summer temperature, and high density residential land within a 1 km window, windspeed, topex and distance to sea. These highlight the important role of topography and meteorological data in influencing, and thus estimating, O_3 concentrations, together with local sources of O_3 precursors (especially NOx) which scavenge O_3 from the surrounding area.

Similar conclusions were drawn by Beelen et al. (2009). Altitude, major road density, high density residential land within 5 km, distance to sea and meteorological factors were again the main predictors in the rural model. In the urban model (comprising both global and urban models), altitude, distance to sea, and meteorological factors (in the global model) and high density residential land within 1km (urban model) were the main predictors.

8.4.2 The space-time models (base and full)

The space-time model in this thesis was developed at two levels of analysis: as a relatively simple 'base model' and a more sophisticated 'full model'. The base model comprises the LUR and the trigonometric functions, weighted by the site-type probability. This base model only explains the systematic variation in O_3 concentration, but can therefore be applied more readily because it needs no locally-specific time-varying data (e.g. on meteorology). As the results showed, the model is most reliable at averaging times of weeks and greater (see Section 7.1). The full model also includes data on meteorology, and explains the systematic and some proportion of the unsystematic variation in the data. Model development and application is more demanding, however, because data are required on local meteorological conditions, and the model needs to be calibrated using regression analysis against data from local monitoring sites. This full model provides more reliable estimates of

exposure than the base model, especially for short averaging times (e.g. days). This section addresses the question of how well these-space time models perform with regard to models and methods used in other comparable studies.

As outlined in section 2.2.3 of this thesis, most previous temporal or space-time models of O_3 concentrations have been based on trying to simulate mathematically the chemical processes involved in O_3 formation and destruction, and the physical processes involved in dispersion. Dispersion models for O_3 comprise three main components:

- A meteorological model (or pre-processed meteorological data) describing the behaviour of the atmosphere;
- A chemistry-transport model, describing formation, destruction and transport processes;
- An emission model (or pre-processed emission data), describing the release of precursor species (including especially NOx and VOCs) from both anthropogenic and natural sources.

One of the most widely-used dispersion models in recent years is CMAQ (Shi et al., 2012, Chemel et al., 2010, Tong and Mauzerall, 2006, Sokhi et al., 2006). CMAQ is designed to operate at the local, regional, and global scales, and can be applied at a resolution as fine as 1 km. This provides a useful benchmark against which to evaluate the performance of any alternative model. Results from applying different versions of CMAQ have indicated some of its limitations. In particular, it is evident that while it can predict average concentrations (i.e. the trend in the data) with some degree of reliability, it is not able to predict extremes of either high or low concentrations (Sokhi et al. 2006, Tong and Mauzerall, 2006). In this way, it parallels the results obtained here.

The optimum way to test model performance is by comparing the model predictions against observed concentrations at an independent set of O_3 monitoring sites and assessing the accuracy of the predictions by using performance measures such as R^2, RMSE, and fractional bias. This is often not done, or is carried out in different ways, which makes it difficult to compare the performance of different models. Other studies, in particular dispersion models, often use slightly different performance measures, notably the Normalised Mean Square Error (NMSE) and the Fractional bias (FB):

$$\text{NMSE}=\frac{(\overline{\text{Co}-\text{Cp}})^2}{(\overline{\text{Co}*\text{Cp}})} \quad ; \text{FB}=\frac{(\overline{\text{Co}}-\overline{\text{Cp}})}{0.5\,(\overline{\text{Co}}+\overline{\text{Cp}})}$$

where $\mathrm{Co}, \mathrm{Cp}, \bar{\mathrm{C}}$ respectively represent the observed, predicted, and average values for the whole dataset. All these metrics quantify the error in the modelled concentration compered to observed concentrations.

A perfect model (i.e. an exact match between the observed and predicted concentrations) would give FB= zero and NMSE=1, therefore, as NMSE becomes greater than 1 it can be concluded that the distribution is not normal but is closer to log-normal (e.g., many low values and a few large values), (Chang and Hanna, 2004). Hanna and Chang (2012) set the threshold for acceptance of NMSE ≤ 3 (i.e. the random scatter ≤ 1.7 times the mean) for rural areas; and ≤ 6 (i.e. the random scatter ≤ 2.6 times the mean) for urban areas. FB=0 indicates no difference between observed and predicted concentrations, FB>0 indicates underestimation in predicted concentrations and FB<0 indicates overestimation. For the purpose of comparison, these additional performance metrics have been calculated for the models developed in this thesis (both base and full models; Table 8.1).

Table 8.1 Performance of the base and full models evaluated by NMSE and FB metrics

Model	Averaged time	NMSE	FB
Base model	Hourly	0.21	0.004
	Daily	0.12	0.002
	Weekly	0.07	0.005
	Monthly	0.04	0.003
The Netherlands full model	Daily	0.12	0.002
The Rome full model	Daily	0.09	0.03

Overall, these statistics reveal a good agreement with the acceptable ranges for the performance statistics: a low fractional bias (0.002 – 0.03) and a very close conformity between average predicted and observed concentrations (as shown by the range of NMSE). Hour by hour comparisons of concentrations in the base mode give a NMSE of 0.21, but this value decreases with the temporal aggregation of the concentrations. The predicted concentrations tend to be lower than the observed concentrations as result of underestimation in predicted concentrations, indicated by the positive value of the FB. Day by day comparisons of prediction and observed concentrations for the full show models in the Netherlands and Rome also tend to slightly underestimate the observed concentrations.

Several other studies have applied and tested dispersion models for O3 and thus provide a basis for comparison. Sokhi et al. (2006), for example, used the CMAQ dispersion model to predict hourly O_3 concentrations for two periods of five days in July and August 2002 in the City of London at a 1km

resolution (described in Subsection 2.2.2.1). Model performance was evaluated at nine independent monitoring stations. Predicted values showed similar trends to the pattern of observed values but the model was unable to predict the extremes (Sokhi et al. 2006). For the O_3 predictions in August, Sokhi et al reported a NMSE of 0.40 and FB of 0.13. In comparison, the base model used here (without any local meteorological data), which also predicts on an hourly basis, gave values for NMSE and FB of 0.21 and 0.004, respectively, suggesting a relatively good performance (the more so because estimates were for several years). It is also notable that the CMAQ model failed to predict the extreme concentrations well, as is the experience here. This implies that the problem is due to factors not considered in either of the models, or simply that the noise in the O_3 concentration data makes it difficult to predict extremes.

Shi et al (2012) also used an improved version of CMAQ (with additional satellite data to enhance the meteorology component of the model) to predict daily O_3 concentrations across Southwest USA in June and July 2006 for a 36km grid. Model performance statistics was reported as R = 0.62. The study area covered by Shi et al is comparable to Western Europe in size but the model in this thesis has a much higher resolution and reports a higher R (0.73 for the base model at a daily level of aggregation). At the city level, likewise, the model developed here seems to perform well compared to CMAQ. Shi et al (2012) ran the latter model for the city of Phoenix in Arizona, and reported an R of 0.76. The full space-time model was applied in this thesis to the city of Rome, and gave R = 0.81; the model developed here thus performs marginally better.

Gariazzo et al. (2007) used a flexible air quality regional model (FARM) to predict daily O_3 and NO_2 concentrations in Rome during three episodes (each five days long) in June and July of 2005 and January 2006 at 1km resolution. Performance statistics were reported for each of the three periods and gave a range of FB from -0.28 to 0.11 and NMSE from 0.21 to 0.91. For this thesis the FB for the model run in Rome was 0.03 and NMSE 0.09.

While comparisons have been made with some dispersion models, there are few space-time models that have studied O_3. One methodological study discussed the use of hierarchical Bayesian modelling of O_3 in three Midwestern states of America (Illinois, Indiana and Ohio). This was carried out to predict the daily maximum 8 hour average concentrations of O_3 over 150 days between May and September in 2006 and used 105 sites for model building and 12 validation sites (DOU et al., 2010). The author compared this method to the CMAQ dispersion model and found it to provide more accurate predictions of annual highest daily maximum 8 hour average O_3 concentrations than did

CMAQ(Sahu, 2011) . The approach used, however, requires a considerable computational effort (Arima et al., 2012).

Compared to other, similar space time models of daily O_3 concentration, the results in this thesis are moderately successful. Bloomfield et al. (1996), for example, modelled daily concentrations between April and October over 1981-1991 in Chicago, USA using a complex high order polynomial function, and allowing for interaction between the meteorological factors (producing a model with 20 terms), obtaining R^2= 0.80 and RMSE=16.4 $\mu g/m^3$. Another study, modelling the daily domain peak of O_3 concentration in Jefferson county, USA, between May and September over the period 1993 – 1996, using polynomial transformations of meteorological variables, which gave R^2= 0.70 and a standard error of 24.2$\mu g/m^3$ (Hubbard and Cobourn, 1998). This suggests that the models might be further improved. In doing so, however, there is a tendency to fit the models more tightly to the data, but to lose some degree of generalizability. Whether this would improve prediction at unsampled locations is thus debatable, and without independent validation (i.e. against an unbiased sample of locations, not used in model-building), the model performance cannot be determined.

It is nevertheless clear that even the full model leaves a substantial proportion of day-to-day variations in concentrations. In particular, the model tends to smooth out the predicted concentrations and to miss the extreme values. In part this may be because the model lacks information on important influences at this level – most notably day-to-day variations in emissions, or the effects of wind direction (which can alter the quantities of precursor pollutants transported from other areas). It may also be due to non-linearity in the relationship between O_3 concentration and the meteorological factors. These were sought, by using various transformations of the meteorological variables, but none of these improved model performance.

Variations in the performance of the full model are also evident between different monitoring sites. In the Netherlands these reflected a general trend in performance from poorer in the north and west (and nearer the coast) to better to the south and east, and inland. Reasons for this are not clear, though it may reflect inadequacies in the meteorological variables used. For example, wind direction (which may have substantial effects on O_3 transport into or out of the area) was not used in these models; nor was any account taken of atmospheric stability and the mixing properties of the lower boundary layer, which likewise affect O_3 removal and build up (Zhang et al., 2012). Incorporation of data on these could help to improve the full model.

Overall, therefore, the model developed here compares well in terms of performance with other, dispersion and space-time models. One advantage of the model developed here is its simplicity in

term of data requirements. Use of dispersion models, by comparison, is often limited by their demanding input requirements and the computational processing time they require. It is particularly important to recognise that a dispersion model requires emission data. Detailed emission inventories are rare, and generalised emissions data (e.g. at regional level) are of little use, for they undermine the accuracy of the dispersion modelling. For this reason, emissions data usually have to be modelled, using data on socio-economic activities, traffic movements, energy usage etc. Typically, these data are also highly aggregated, so some form of spatial (and temporal) disaggregation is required. Thus, dispersion modelling tends to rely on a series of prior models, with the inevitable risk that substantial (and often unseen) error in the input data. This was clearly demonstrated by the work to develop emissions data for the EU, at 1 km level, as part of the APMoSPHERE study (http://www.apmosphere.org). While reasonably good estimates of annual average emissions could be obtained for NOx, estimates of other pollutants (including PM) were poor, largely because important elements in the emission process were not well quantified (e.g. source distribution or intensity).

Perhaps for these reasons, dispersion models have rarely been used over both an area and a time period comparable to those used here. On the other hand, the similarities in results between these more complex and detailed dispersion models and the approach used here suggests that nothing is being lost with the reduced data input. Conversely, it suggests that significant improvements in O_3 modelling beyond those achieved here will come only from an even more detailed understanding of the processes of O_3 formation, dispersion and production – and is thus likely to require even more detailed and reliable data, and even more powerful computer processing facilities.

8.5 Sources of variability in O_3 concentrations in Europe

Each of the three components of the full model – the LUR model, the trigonometric functions and the meteorological variables – explains a different component of the variation in observed O_3 concentrations. Results from pilot study (Table 4.3) – based on hourly O_3 measurements, for 1253 monitoring stations across Europe from March 2001 to February 2002 – suggested that ca. 14% of the overall variation was spatial and 28% temporal. Of the temporal component, the majority relates to season (18%), while diurnal variation (14%) makes up most of the rest: the hebdomadal variation was thus small. On the basis of these estimates, ca. 58% of the variation is left unexplained (error), and thus comprises noise.

Subsequent analysis, using six years of data for 1211 monitoring stations in building the space-time model, broadly confirmed these proportions, though succeeded in explaining a larger proportion of the overall variation. The LUR model explained 67% of the spatial variation in the long-term mean, with a standard error of 7.6 $\mu g/m^3$, better than what was suggested in the pilot study (Table 4.2). The temporal model, built using the trigonometric functions, explained ca. 42% of the remaining variation in the hourly data. Of this, ca. 63% related to seasonal variation and 36% to diurnal variation. In the Dutch and Rome case studies, including the meteorological factors explained about 10% additional variation, leaving about 47% or the variation unexplained variation in the Netherlands and 35% in Rome at the hourly level. The unexplained variation at the end of the modelling is thus somewhat less than that suggested by the initial pilot study (63% in the Netherlands and 49% in Italy).

The different components of the time functions vary in their importance with site type environmental characteristics. Systematic variation tends to be lower in topographically exposed areas, probably because the vertical distribution of O_3 varies more randomly, due to short-term and local variations in transport, deposition and chemical reactions. In most site types, however, seasonal variation is strong, and is often well specified by the models. This is largely because seasonal differences in O_3 concentrations are driven mainly by the seasonal trend in temperature and solar radiation, which show a clear systematic pattern. In contrast, the hebdomadal patterns are very weak, and contribute little to overall variability. Diurnal patterns tend to be relatively clear, but vary between different site types and are not well characterised by the models, leaving substantial variation unexplained. This appears to be because patterns vary both from day to day and from site to site, depending on local, short-term factors such as emissions and weather. Incorporating meteorological data improves model performance by providing information on some of these short-term variations, but because these data were only available at a daily level, they could not be used to model the hour-to-hour variations.

Overall, these results suggest that obtaining better data on temporal variations in emissions could help to reduce the uncertainty in the modelling. Emissions inventories are widely available in Europe, both at a broad, continental scale, and more locally (e.g. for individual countries and cities) (Davison et al., 2011), but these rarely provide data for timescales less than a year – and even those may be highly generalised. For the most part, therefore, data are not available at the spatial and temporal scale needed, and in any case are subject to their own uncertainties. The best that can probably be achieved, therefore, is to estimate the statistical distribution of the variability in emissions, in different site types. One way of doing this is by using Bayesian Monte Carlo techniques.

Bayesian Monte Carlo techniques are used to improve the estimation of output uncertainty in modelling data by adding in an uncertainty distribution that reflects the noise components of the data (Wang et al., 2012). For example, in water quality modelling this approach reduced the uncertainty in the model by 72% (Dilks et al., 1992). The main problem with this approach is that it needs a good understanding of the technique and suitable computational facilities (Dubus et al., 2003, Dilks et al., 1992).

8.6 Uncertainties and limitations

Any of air pollution models are liable to uncertainties. These typically originate in two main sources of error:

1. in the quality and completeness of the input data (measurement or sampling error);
2. in the structure and parameterisation of the models (modelling error).

8.6.1 Measurement or sampling error

Regarding the quality of the input data, the errors of most importance are those associated with the AIRBASE data, since these were fundamental to every step of the modelling, and also provided the data used to assess model performance. Measurement errors inevitably occur in the instrumentation of air pollution – both in terms of calibration and operation of the instruments. As a consequence, data were often missing, and errors were apparent (e.g. the occurrence of negative concentrations in the observed concentration data). Obvious errors of this type were relatively easily spotted and allowed for. More difficult to deal with were the hidden uncertainties, such as false (though plausible) readings, or transcription errors – and even more importantly the sampling error in the data. The networks from which the observed concentrations were obtained were both unevenly distributed and sparse, meaning that not every area was well-represented. Inevitably, this biases the model towards better represented areas and means that prediction error may vary geographically.

Differences also exist in the way the air pollution monitoring networks have been set up in different countries, in terms of the distribution and sampling density of sites which may contribute some of the uncertainty to the model. In addition, the distribution of sites within countries may differ. There are areas where sites are clustered – for example, in some metropolitan cities such as London

(Figure 3.5). Some countries have intensive monitoring networks with a comparatively large number of sites, such as Austria, Germany and Spain, while in others they are widely distributed and sparse (e.g. Ireland). The model will inevitably be more heavily weighted toward countries that provide a larger number of sites, and towards areas where sites are clustered, and may not represent areas with less representative monitoring sites as well. This variation in the geographical distribution of errors was apparent in the sensitivity analysis for LUR, with errors varying between location, topography and urban/rural areas as discussed in Section 5.4.2.5.

It should also be noted that the site type classification used to select the original AIRBASE sites is not specific to individual pollutants, and is not geared to O_3. Even where countries follow the classification rigorously, therefore, the networks do not necessarily provide representative sampling of O_3. Biases in the data are therefore likely, which inevitably affect both the calibration of the model and the results of validation. This highlights the importance of establishing an objective, pollutant-specific approach in classifying AIRBASE sites across the EU. This would not only help to make the network as a whole more representative and consistent, but would also potentially enable countries to rationalise their networks and ensure that they were efficient.

8.6.2 Modelling error

A further source of uncertainty comes from the structure and parameters defined for the models. The limitation of the predictors in representing the processes of interest also has to be recognised. One of the main limitations is that all the variables used in the analysis relate mainly to local factors; little explicit account is taken of more distant effects, for example as a result of long-range transport of O_3 (operating over distances of tens to hundreds of kilometres and timescales of several days). Long-range transport of air pollution in general, causes frequent episodic pollution events (Monks et al., 2009). Europe is affected by long range transport of O_3 from North America (Wild and Akimoto, 2001 cited in Monks et al., 2009), as well as extensive cross-border transfers between countries. The result of this is likely to be in the errors evident in the models – notably in the occurrence of several successive days of under-estimation, which probably reflect pollution episodes due to long-range transport. Attempts were made, on an exploratory basis, to model these by adding post-hoc functions to the temporal models, but these proved to be of limited value (Section 6.5.1).

One way of including these regional effects of long range transport in the model might be by incorporating satellite data on atmospheric O_3. A number of studies have used information from satellite-based monitoring, usually together with modelling, to estimate and map regional and global

tropospheric O_3 (Ziemke et al., 2006, Martin et al., 2002, Fishman and Balok, 1999, Chance et al., 1997). The data provided by satellites includes not only measurements of O_3 concentrations in the atmospheric column, but also other relevant data such as regional-scale meteorological conditions that might lead to pollution episodes (Fishman and Balok, 1999). The coarse resolution both spatially and temporally of satellite data, and the fact that they are restricted to clear, non-cloudy days, limits their use. However, the fact that the data are updated regularly means that they can provide valuable information for modelling. One way of using them, for example, is as regional-scale data for use in the LUR models. Another is to use them to recalibrate the LUR models to reflect changes in O_3 levels and distribution. In this case they can be seen as a temporally varying canvas on which the finer detail of the LUR model is printed. While the detailed patterns probably vary little over time, because they are driven largely by a fixed pattern of local emission sources, the whole canvas is warped and changed as regional patterns of O_3 vary in response to regional-scale meteorology and long-distance transfers of O_3. In this way they might greatly improve the temporal models.

Another limitation inherent in the model is the use of proxies. Land cover data, for example, have been used as a proxy for emissions. As already noted, emission data at the local scale are not available and the data that do exist have high levels of uncertainty because of their level of aggregation (Davison et al., 2011). Many emission inventories are also, themselves, products of models that have down-scaled national estimates to a regional or local scale. The extent to which these proxies are able to reflect directly the processes or effects they are intended to (e.g. emission sources/intensity) is limited and variable. For instance, for O_3 prediction VOCs are one of the most important precursors. Biogenic VOCs, for example, have been shown to play an important role in controlling O_3 formation downwind of power plant plumes (Ryerson et al., 2001). At a global level, VOC emissions are greater from biogenic sources (ca. 1150 Tg C /year) (Guenther et al., 1995) than from anthropogenic (ca.150 Tg C /year) (Müller, 1992). At the regional level, there is more variability in the relative emission rates: in Europe the national emissions can be dominated by either source (biogenic or anthropogenic), depending on the land use features of different countries (Simpson et al., 1999, Simpson et al., 1995). Overall, however, biogenic VOCs are more active by 2-3 times than anthropogenic VOCs, as mentioned in Section 3.2.3.1. Direct data on local VOC emissions are not available, so in the LUR model forest land cover classes were used as a proxy, due to the fact that active biogenic VOCs (especially isoprene and monoterpenes) are emitted from forest plants. Forest land cover was estimated by combining three land cover classes; broad-leaved forest, coniferous forest and mixed forest. Maucha and Büttner (2005) emphasised that these three classes of forest

land cover were accurate between 85% and 90% on average at the European level when compared with the LUCAS survey (a ground-sampling-based survey in 18 European countries) for 2001 and 2002. On the one hand, this suggests that the forest land is probably well defined in the land cover data, though small changes in the distribution of forest will have occurred between 2001/2, when the data were compiled, and the years covered in this thesis study. But more important is the lack of thematic detail in these data. Areas mapped as forest inevitably vary substantially in species composition, age, density of the trees and canopy cover, rate of growth and transpiration rates – all of which affect VOC emissions conditions (Schurgers et al., 2009).

Another limitation in using the land cover data as an emission proxy in the spatial model and to predict site types is that they are known to suffer from a range of uncertainties. The accuracy and classifications for land cover are not wholly consistent between countries and only limited assessments of a few habitat types have been carried out to evaluate the data (Boresjo Bronge and Naslund-landmark, 2002, Kennedy and Bertolo, 2002, Martin de Santa Olalla Manas et al., 2003cited in (Waser and Schwarz, 2006). The land cover data are also to some degree generalised. All features in the original vector database were classified and digitised based on satellite images with 100 m positional accuracy, which means that in this study the data are being used at the very limits of their detection: at a 100m resolution, therefore, there is a significant risk of misclassification. To some extent, this was reduced by aggregating classes, but in doing so, some information was lost, and while aggregation improves the class-level consistency (i.e. it reduces the prevalence of areas assigned to the wrong class), it increases the within-class uncertainty (because each class represents a wider range of land cover types). Nevertheless, CORINE land cover is the most widely used and consistent land cover data for Europe and reducing some of the uncertainty resulting from its use is beyond the scope of this research.

Similar uncertainties result from the use of road data (road length by road type) as a transport emission proxy. Although the road data were of a high spatial resolution, there were obvious differences in the way roads had been classified in different countries. To reduce these, the initial original seven FRC classes were reduced to three classes representing major, secondary and local roads. This inevitably resulted in some loss of information. Moreover, the relationship between road types is inevitably rather weak and variable. The same length of one road type in a rural and urban area, for example, might well have different traffic intensities; in the model, both would be applied the same coefficient regardless of the real impact on the O_3 concentration. This highlights that, as scavenging is crucial in determining O_3 concentration, it is important to include traffic flow rates in O_3 models. Currently, however, data are not available across Europe.

The meteorological data used in both the spatial model and the full space–time model could have introduced some uncertainty into models. In some cases, this may have altered the way in which the meteorological variables operate in the models. As mentioned in Section 5.4.3, the meteorological data used in the LUR model had been rescaled using IDW from 40Km resolution to 100m, the data were thus capable only of showing broad, regional-scale effects.

In this study, as has been noted, windspeed was consistently, positively associated with O_3 (section 8.3). Even when the O_3 data were analysed by month, using data from the Netherland, the correlation remained strongly positive (R=0.60) for most months and showed only marginally negative correlations (R=-0.07) for the summer months (June – August). This may indicate that wind speed is associated with vertical mixing, so that O_3 formed near the ground can be mixed into the upper levels during the day, and O_3 trapped high up during the night is brought down to the ground level the following day. When vertical mixing is high, O_3 may thus be entrained from the O_3 rich layer in the higher atmosphere (Kim et al., 2007, Rao et al., 2003). This highlights the lack of data on either mixing height or atmospheric stability in the models. Data on wind direction were not included and without this, it is impossible to determine the physical relationship between monitoring sites and their source areas (either of O_3 or of scavengers), and thus to deduce the effects of local or regional transport of O_3 on concentrations.

In conclusion, O_3 is a complex secondary pollutant and as a result presents a large challenge modelling, which thereby leads to the limitations mentioned here. The inclusion of predictors as a proxy for any mechanism in the processes of generating O_3 therefore has to be very well understood, not only for the predictor itself but also its relationship with other predictors in the same model. Despite these various limitations, the space-model developed here was shown to perform as well as, and in some cases better than, previous dispersion models (section 8.4.1). Based on these results, the model thus has potential value as a basis for exposure assessment. This is discussed with limitations also in consideration, in the following section.

8.7 Strengths and application

Exposure to O_3 leads to harmful consequence on human health, as addressed in Section 2.1.4, in terms of both short and long term risks of respiratory and cardiovascular morbidity and mortality (WHO, 2008; UNECE, 2008). There is also growing evidence that exposure to O_3 during pregnancy might be associated with adverse birth outcomes (Le et al., 2012, Hansen et al., 2009, Salam et al.,

2005). The key challenge in exposure assessment has been in depicting the spatial and temporal variations in air pollution (here O_3 concentrations) at an accuracy sufficient to allow detection of these effects.

This space-time model can be used for exposure assessment in a range of different contexts. In epidemiology, there is potential to use the approach in a range of different study designs. In this section, the potential for applying the models for exposure assessment will be discussed, emphasizing the strength and limitations.

8.7.1 Strengths of the spatial model

As noted, the major motivation for developing this model of O_3 concentrations is as an aid to exposure assessment in risk analyses, health impact assessments and epidemiological studies. Where these concern the chronic effects of air pollution, the spatial model alone is of potential value, for it provides estimates of the annual (or long term) average O_3 concentration, at a small area resolution (and close to that representing an individual home). It is therefore instructive to determine what level of improvement, if any, is given by this model compared to other, more traditional ways of estimating exposures.

The most common approach for exposure assessment in epidemiology has generally been to estimate exposures directly on the basis of the monitored data – by simply attributing each home to its nearest monitoring site. This approach assumes (though rarely states) that the pollution surface is slab-like, with flat areas (in terms of pollutant concentration) around each monitoring site, and vertical disjunctions between them. The extent to which this provides an accurate estimate of exposures can be assessed by comparing concentrations at each monitoring site with that at its nearest neighbour: if the approach is valid, the association should be strong, and the prediction error small. To explore this, the correlation was assessed between the long term average (from 2001-2007) for O_3 concentrations at each monitoring site and its nearest neighbour for all sites in the AIRBASE data set within the study area. This was done for the training dataset (979 sites) of monitoring stations. Results gave R^2=0.14, and RMSE=12.13ug/m^3. The analysis was repeated, stratified by country, but this gave no improvement in the results (Appendix B, Section V, Table B.5). Figure 8.3 shows the scatterplot between the two estimates and the 1:1 line. As can be seen below while there is a relatively high proportion of estimates cluster around the 1:1 line, there is considerable scatter around this cluster. In general, estimates from the nearest neighbouring site

tend to under-estimate actual concentrations at low to moderate levels but to over-estimate them at higher concentrations.

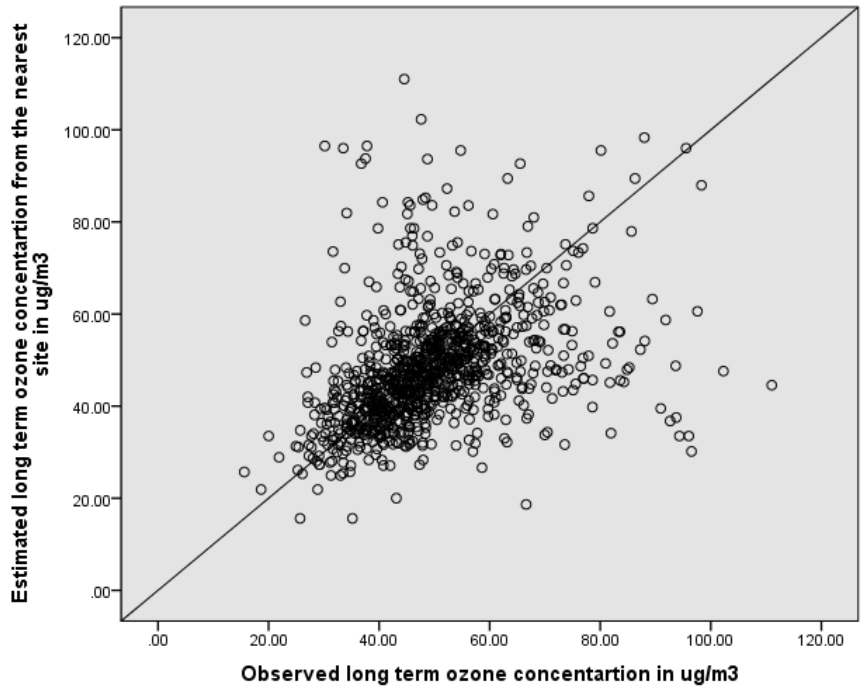


Figure 8.3 Scatterplot between the observed long term O_3 concentrations and the estimated concentration from the nearest monitoring sites using the training dataset (979 sites)

The results thus suggest that estimating the exposure from the nearest site or by a simple interpolation method based on local sites is likely to be misleading, and may not represent exposures with any degree of reliability. The alternative, of modelling the pollution surface by taking into account the characteristics of surrounding environment, as is done with LUR, is likely to be more effective.

The further assumption is often made that neighbouring sites are more closely correlated than those further away. Further exploratory analysis was therefore done by categorising the distance between nearest monitoring sites into six categories as shown in Table 8.2, and calculating the correlations within each distance class. The relationship showed a variation with distance, the mean difference between each pair of monitoring sites increasing, and the overall R^2 falling, as distance increases.

There is also a suggestion in the data of a threshold distance of less than 1Km (range between 167 m to 993 m), within which the differences are small. This suggests that data can be safely extrapolated over a distance of up to 1 Km, but beyond that errors increase and estimates of exposure become progressively less reliable. The same finding was reported by Moral García et al. (2008), as mentioned in Section 5.1.

Table 8.2 Correlation between nearest monitoring sites in different distance

Distance	No. of sites	R	Adj. R^2	RMSE
<1Km	19	0.83	0.67	6.53
1 -3Km	122	0.67	0.44	7.12
3 - 6 Km	149	0.47	0.21	10.18
6 - 10 Km	145	0.39	0.15	10.54
10 -20 km	177	0.37	0.14	13.57
>20 km	367	0.13	0.01	12.86

On the basis of the results presented here, therefore, it can be argued that the LUR model can predict the spatial pattern of long term mean O_3 concentration across Western Europe at a 100m grid resolution to an acceptable level of accuracy. The overall performance of the LUR model is comparable to the few previous attempts to model O_3 at the continental scale, with R^2~0.7 and RMSE~7.0μg/m^3. Performance of the LUR model, however, shows some weakness in representing specific types of areas, and performs better for example at rural than it does at urban sites, at higher altitude rather than low altitude sites, and in some countries compared to others (Section 5.4.2). It also needs to be recognised that the model should not be used outside this study area or time period without further validation and calibration.

8.7.2 Strengths of the space-time model

The modelling approach developed here might help to overcome the limitations of the traditional approaches of exposure assessment (such as use of the nearest monitoring site) discussed in Section 7.4 Risk assessment, health impact assessment and epidemiological studies, however, involve the analysis of a wide variety of exposures, ranging from the population level to the individual. Exposures may likewise need to be assessed over averaging periods from a few hours or days, in the case of studies of acute health effects (e.g. using time series designs) to years where the health effects have long latency times or where accumulated exposures are of interest. The question thus

arises: what spatial and temporal precision is required for different types of application, and to what extent does the model presented here meet these needs?

The temporal precision of exposure assessment is important but has several facets. It relates to both the duration of the critical exposures and to the timing both in absolute terms and relative to manifestation of the health effect. The current literature demonstrates that there are varying critical time windows between O_3 exposure and health outcomes, depending on the disease of interest. Many epidemiological studies, as well as risk assessments, of O_3 have focused on acute effects. Examples include not only studies of short-term respiratory reactions amongst susceptible individuals such as asthma sufferers (Fusco et al., 2001, Ponce de Leon et al., 1996) but also admission to hospital for cardiovascular diseases (Park et al., 2005, Brook et al., 2002b), as discussed in Section 2.1.4.2. In these cases, the critical exposure window is usually assumed to be the same day or one to three days before manifestation of the disease, and exposures are often measured as the average daily, or maximum (e.g. 8-hourly), O_3 concentration. There is, however, some evidence that even shorter exposure durations may be important in terms of health. One study, for example, found an association between the one hour lag of maximum O_3 concentration with arrhythmia (Rich et al., 2005) and another found maximum daily O_3 concentrations to be associated with an increased risk of death from respiratory diseases (Jerrett et al., 2009).

Rather less attention has been given to the effects of long-term exposures in relation to disease initiation. In the case of rhinitis, however, effects of exposure to O_3 may take a year to develop, while for asthma it may take as much as fifteen years (Section 2.1.4.1, Table 2.5). Adverse birth outcomes have also been associated with O_3 exposure over different time periods ranging from weeks to months to trimesters or longer. For instance, an increased risk of heart defects was associated with an increase of $10\mu g/m^3$ in the 8 hour O_3 concentration, averaged over weeks 3-8 of pregnancy (Hansen et al., 2009). Low birth weight was found to be associated with increases in exposure to O_3 in pregnancy, measured using 24 hourly averages and the average daytime concentration (between 10.00-18.00 hours) over the entire pregnancy (Salam et al., 2005).

These examples emphasise that the critical time window between O_3 exposure and health outcome is dependent on the outcome of interest. They also suggest that it is not only the average concentration over an extended period, but also in some cases the peak exposure within a period of a day, or even an hour, that might be critical in causing health effects. It is further evident that in many cases the real causal period of exposure is unknown, either because exposure data have not

been available for a suitably precisely defined period, or because strong temporal autocorrelation between the measurements makes it difficult to disentangle the effects of one period from another.

In the case of many time series studies (including some cohort studies) the spatial variability in O_3 concentrations has been addressed by applying monitored concentrations either at one central site or by averaging across a number of sites in the study area. This is a limited approach as large numbers of people are assigned the same exposure for any one day, even though they may live in areas with very different pollution levels. Two assumptions are thus made: that the absolute change in exposure from one day to another is the same for everyone (e.g. when the monitoring station shows a 10ug/ug^3 increase in concentrations from the previous day, this is what everyone experiences); and that the exposure-response relationship is linear, so that the effect of a 10ug/m^3 change in exposure is the same for everyone, regardless of the underlying concentration. If either of these assumptions is invalid, study sensitivity is likely to be lost due to the dilution of any relationship with health by errors in the exposure assessment, and the estimated relationship may in some cases be biased. To avoid these dangers, and to provide a basis for detecting more reliably the shape and slope of the exposure-response curve, as well as the critical periods of exposure, requires exposure data that are more finely resolved, both spatially and temporally. A good O_3 exposure model should thus have the temporal precision to estimate individual-level exposures over periods of a day or less, and – in the case of peak exposures – for durations of only an hour or so. The results presented here suggest that the models developed in this thesis can provide this level of spatial and temporal specificity.

It has often been argued that temporal correlations between different sites at the daily level are high; this is the justification for averaging the data from different sites in time series studies. So the mean concentration at nearby sites might vary, but the concentrations will tend to rise and fall together. If so this would give some justification for the time series approach. If not, then assessing exposures in time series studies would benefit from a different approach. A number of time series studies have explored temporal correlations between O_3 monitoring sites and these typically report a correlation of R= ca. 0.45 (Ballester et al., 2001, Fusco et al., 2001, Morgan et al., 1998, Medina et al., 1997). To explore this issue, the correlation between the daily averages of the five sites in Rome city with their nearest neighbour (where the nearest distance was about 6 Km) were calculated. The average R^2 was 0.34 with RMSE= 19.80 μg/m^3.

These results can be compared with those from the base model. As noted this was developed and tested for a range of averaging periods, from hourly to monthly. At the hourly level the performance

was only moderate, explaining only on average 37% of the variation and with RMSE = 22 $\mu g/m^3$. Even so, this is similar to the daily average for the inter-site correlation in Rome. Moreover, model performance improved as the data were aggregated to daily, weekly and monthly level, and from the daily level onwards further improvement was possible by including meteorological variables, as part of the full model. For the thirty five monitoring sites in the Netherlands and Rome, this gave an average R^2 of 0.59; in Rome, where the underlying contrasts in daily exposure were higher, the average R^2 was 0.69. This is comparable with the levels of accuracy usually achieved by exposure models for much longer averaging times – e.g. land use regression models or dispersion models of annual average concentrations. It is also markedly better than is likely to be achieved by using data from the nearest monitoring site.

The real potential of the time-space model developed here, however, is likely to come in its use in cohort studies. By their very nature, these require space-time exposure estimates, because they cover a large number of individuals who live in different places and who need to be followed up for long periods of time. Crucial exposure periods may vary from a matter of days for many respiratory effects through to months for pregnancy outcomes, and to many years for long latency diseases such as many cancers. This model allows flexible time- and location-specific exposure estimates to be generated for individuals over a wide area.

The specificity of these estimates has a further advantage. Characterising and maximising the exposure contrast within a cohort or study population is an important issue. If the contrast in exposures between individuals is lacking, the impact of the agent (in this study O_3 concentration) may not be accurately assessed, and may be missed. Avoiding exposure averaging reduces dilution of the relationship with health, and helps to retain study power (Nuckols et al., 2004). The same ability to estimate local variations in concentrations across an area also improves exposure estimates in other study designs, such as case-control studies (i.e. where outcomes are often compared in high and low exposed groups), or in panel studies where a small group of individuals with a known risk need to be followed up.

In the same way, exposure estimation using the approach developed here may help to separate of effects of determinants of health which are otherwise confounded. The 2003 heat wave in Europe, for example, involved an almost contemporaneous increase in temperature, O_3 and PM (Alebić-Juretić et al., 2007, Poumadère et al., 2005) and It was associated with substantial increases in premature deaths. Attributing these deaths to the different potential causal factors, however, was difficult, because of the spatial and temporal uncertainty of exposure estimates. Stedman (2004)

estimated that, in the UK, there was an excess of 423 premature deaths associated with O_3 exposures, and 769 associated with PM_{10} exposures. In the Netherlands, Fischer et al. (2004) reported 400 excess deaths attributed to O_3 exposure and 600 deaths attributed to PM_{10}. In France, however, Vandentorren et al. (2004) attribute the cause of the majority of deaths to the effects of heat and home condition. In situations such as this, therefore, there is a need for a standard way of assessing exposures across a large study population, at high spatial and temporal resolution in order to separate out the effects of these different potential risk factors, on the basis of subtle differences in their timing and geographic distribution. The approach developed here offers such an approach for O_3 exposures.

Another potential application relates to the current threshold for O_3 (120μg/m^3 for maximum 8-hours average and 180μg/m^3 1-hour average), which has been set largely on the basis of chamber studies. As WHO (2008) reported, no such threshold has been detected at population level in epidemiological studies. In part, this may be because no epidemiological study has so far used a sufficiently large population, with individual level exposure data, capable of distinguishing a threshold. Monks et al. (2009) argue that if the relationship between the health outcome (response) and O_3 concentrations (exposure) is linear, then the total annual health impact is relative to the annual O_3 mean concentrations. This would imply that O_3 concentrations even at or below the policy thresholds will continue to have substantial health effects.

As these examples indicate, epidemiological studies and risk assessments increasingly need better methods of exposure estimation, in order to address more subtle and complex research questions, for larger study populations as this model potentially provides. These advances also increasingly mean that exposure assessment needs to be done in a space-time framework.

8.7.3 Application of the models

Both the spatial and space-time models developed in this thesis been provided and applied to number of health studies. The long term estimates from the spatial model were shared with several European colleagues investigating health effects of air pollution. In each case, the estimated concentrations were extracted for the requested cohort participants (by x,y coordinates) to derive exposure estimates on an individual level. This was done for the following studies:

1. GA2LEN study: Global Allergy and Asthma European Network (http://www.ga2len.net/). The long term O_3estimates were sent to Prof. Debbie Jarvis at the National Heart and Lung Institute, Imperial

College. They were used to assess the association of a marker of systemic inflammation, the C-reactive protein (CRP), with long-term exposure to three air pollutants (PM_{10}, NO_2, and O_3). The study areas included are Amsterdam (NL), Brandenburg and Duisburg (DE), Bromley and Southampton (UK), Ghent (BE), Odense (DK), and Palermo, (IT). The result shows that there was no evidence of an association between CRP and the long term concentrations (Ramond, 2012).

2. SETIL study, Italy (Studio sulla Eziologia dei Tumori Infantili Linfoemopoietici): an Italian epidemiological study on the aetiology of childhood leukaemia, lymphoma and neuroblastoma. This work was led by Dr. Forastiere and Dr. Badaloni at the Department of Epidemiology, Lazio Regional Health Service, Rome. The aim of this work was to assess the impact of air pollutants (PM_{10}, $PM_{2.5}$, NO_2, and O_3) on childhood leukaemia. Results showed that there were no associations with these air pollutant. A draft publication, of which I am co-author, is ready for submission (Badaloni, et al., in prep: Occupational Environmental Medicine).

3. PELAGIE cohort in Brittany, France is a prospective birth cohort designed to study the role of environmental pollutants on intrauterine and child development in three Breton districts (Ille-et-Vilaine, Côtes d'Armor, and Finistère). The exposure estimation for 4857 women was sent to Jean-François Vie who is the supervisor of PhD student working on the PELAGIE cohort.

The daily exposure estimates from the space-time models were also shared with three cohorts studies included in the recently funded ESCAPE project but the analysis using the data have not yet been completed.

1. PIAMA in the Netherlands consists of 10,819 pregnant women (as mentioned in section 7.2)
2. ABCD consists of 12,682 pregnant women in the Netherlands (as mentioned in section 7.2).
3. The GASPII birth cohort consists of 713 pregnant women in Rome, Italy (as described in section 7.3).

8.8 Future work

Like any model, that developed has substantial room for improvement. The purpose of improvement might vary – for example, to make its use easier, to make it more readily transportable to other areas, or to reduce the errors in its predictions. It is the last of these which is probably most

important here. The question that has first to be answered, therefore, is what is the cause of these errors in the model?

8.8.1 Enhancing the modelling approach

The approach of creating the time functions, taken here, was based on the principle of using prior knowledge about the factors that were responsible for variations in O_3 concentrations over different time scales (diurnal, hebdomadal, and seasonal), and in different environmental contexts (urban, rural, mountainous, etc). Based on this, a set of time functions was generated of predefined periodicity, and then regression analysis used to estimate weights for each of these, to represent their amplitude. In this context, the question is: do the functions generated in this way adequately reflect the variability in the data?

Improvements would certainly seem possible. Currently, the error (as defined by the RMSE) is ca. 14 $\mu g/m^3$ in the daily predicted concentrations. For average concentrations of ca. 50 $\mu g/m^3$, this represents an error of ca. 28% - more-or-less at the margins of what would usually be tolerated.

To reduce these errors, one approach would be to modify the time functions by allowing both the periodicity and the amplitude to be defined statistically for different data sets. With large data sets such as those used here, this would be time-consuming, because it would involve using a formal grid-search technique, in which each variable is adjusted one increment at a time, whilst keeping all others constant, until the optimum solution is achieved. Related to this, there may be scope to gain some improvement in model performance by allowing for interactions between the systematic patterns at different time-scales – for example, of different diurnal patterns on different days of the week, or in different seasons.

More important, however, might be to improve the assessment of the non-systematic components of temporal variation. Currently, these are modelled only through the use of meteorological variables, using linear models. Non-linear models might be pursued more effectively, as illustrated by the studies of Hubbard and Cobourn (1998) and Bloomfield et al. (1996). As has been recognised, however, non-systematic variation is not only weather-related but also due to changes in emission rate, often in response to weather-related effects. Incorporating data on hourly or daily emission rates of O_3 precursors would almost certainly improve the model and help to obtain better prediction of the extremes values.

The model might also be improved by developing a spatial model for each site type. It was apparent in this analysis that the global model did not perform equally in all site types. For some, therefore, a site-type specific LUR model is likely to provide better results. The reason for not doing it in this thesis was primarily recognition that classification of the site types was itself uncertain, both using HCA and MLOR. To resolve this issue, the probability of site type membership was used to classify unmonitored sites. This would make building specific-site type LUR somewhat complex, for LUR predictions from each model would likewise need to be weighted by their probability of site-type membership. This weighting procedure certainly has an advantage, in that it would help ensure that the resulting modelled surface was not blocky, with major disjunctions between areas of different site-type. Further analysis would thus be merited, exploring this possibility. Nevertheless, since the assignment of site types to the classification (using MLOR) was based on many of the same environmental variables as used in the LUR model, little advantage might actually be achieved.

8.8.2 Enhancing the data

As noted earlier, to help reduce the problem of lack of sparse or uneven coverage and unrepresentatively of the monitoring sites, satellite data could be used as a source of monitoring data. There are now daily maps of O_3 concentrations, at a spatial resolution of 10-20 km across Europe (Ziemke et al., 2011), produced from satellite observations. It is thus possible to combine satellite with monitoring data in order to provide better representation in areas where the ground sites are sparse. In this way, satellite data could be used to reflect the regional variation in the concentrations, and then the model applied to add local detail to this on the basis of land cover and other data. The main reason for the limited resolution of the satellite data is that the sensors measure the entire atmospheric column O_3 rather than surface concentrations. At present therefore, satellite data alone are unable to achieve the spatial resolutions needed for exposure assessment. Harnessed to a space-time model, however, they offer considerable capability.

Linkage of the space-time model to satellite data might also enhance the modelling of temporal variations. One way of achieving this is to use the satellite data as a basis for data assimilation (Vijayaraghavan et al., 2008). This involves continuously recalibrating the model to new satellite data, as they become available. In this way, the model is trained not only to current conditions, but to underlying trends in the data. This helps to improve, also, its ability to predict future developments – and thus enhance its scope for forward extrapolation. For these reasons, research

to link space-time modelling and satellite data for the purpose of exposure assessment needs to be a major priority.

8.9 Conclusion

Key findings in this study are thus as follow:

1- Categorizing the 1211 O_3 monitoring sites across Western Europe produced thirteen site types based on the temporal indicators used. This highlights that the temporal variation of O_3 behaves differently in different site types and this needs to be recognized when exposure is to be estimated.

2- LUR is capable of explaining 67% or more of the spatial variation in long-term average O_3 concentrations across Western Europe, at a very fine resolution (100m) and using readily available data. On the other hand, estimates of average annual concentrations at the 731 training sites from their nearest neighbouring monitoring site gave poor predictions (R^2 = 0.14), and the association declined markedly as the separation distance between the target and source site increased beyond 1 km.

3- Using Fourier analysis to build an hourly temporal model for 13 site types successfully explained, on average, 42% of the temporal variation in the data. Hourly and daily O_3 concentration variations, especially, require additional predictors to represent the non-systematic variation.

4- The base space-time model (i.e. O_3 long term concentration + weighted hourly temporal model) explained 46% of daily O_3 variation. Aggregating the base model to a monthly average increased the percentage of the explained variation to 74%, demonstrating that a large proportion of the temporal variation for these longer time scales is systematic, and can be modeled with relatively simple Fourier functions.

5- The full space-time model was able to predict daily O_3 concentrations over the applied areas (Netherlands and Rome) with a reasonable degree of accuracy, explaining an average of 57% of variation in daily O_3 concentrations, with only moderate uncertainty (RMSE = 14 $\mu g/m^3$).

6- Daily O_3 estimates from the spatial and the space-time model of this thesis in comparison with estimates from the nearest monitoring sites showed weak correlation (R2=0.14 and 0.12 respectively).

7- Space-time modeling, as performed in this thesis, can be used to provide a powerful model, with the strong advantages of combining reduced data requirements (compared to dispersion models) with reasonable computational processing power and time.

In conclusion as epidemiological studies and risk assessment are faced with the increasing challenge of identifying the impacts of O_3 concentrations on human health, against a complex range of confounding factors; better methods of exposure assessment are inevitably needed. These methods need to provide time-varying estimates of exposure at the individual level for large study populations. The type of model developed in this thesis offers one way to face this challenge. In addition it offers a rich field for further research aimed at improving the modeling approach and the data on which it applies, and applying it in future studies of the health risks of O_3.

References

Abdel-Kader, F. H. (2011) Digital soil mapping at pilot sites in the northwest coast of Egypt: A multinomial logistic regression approach. *The Egyptian Journal of Remote Sensing and Space Science,* 14**,** 29-40.

Alebić-Juretić, A., Cvitaš, T., Kezele, N., Klasinc, L., Pehnec, G. & Šorgo, G. (2007) Atmospheric particulate matter and ozone under heat-wave conditions: do they cause an increase of mortality in Croatia? *Bulletin of Environmental Contamination and Toxicology,* 79**,** 468-471.

Anderson , H., Atkinson , R., Peacock , J., Marston , L. & Konstantinou , K. (2005) Meta-analysis of time-series studies and panel studies of Particulate Matter (PM) and Ozone (O3). Copenhagen, World Health Organization

Andersson, C., Langner, J. & BergstrÖM, R. (2007) Interannual variation and trends in air pollution over Europe due to climate variability during 1958–2001 simulated with a regional CTM coupled to the ERA40 reanalysis. *Tellus B,* 59**,** 77-98.

Arima, S., Cretarola, L., Jona Lasinio, G. & Pollice, A. (2012) Bayesian univariate space-time hierarchical model for mapping pollutant concentrations in the municipal area of Taranto. *Statistical Methods & Applications,* 21**,** 75-91.

Armstrong, J. S. (2001) Principles of Forcasting: A Handbook for Researchs and Practitioners. IN J. SCOTT ARMSTRONG, E. (Ed.) *Extrapolation for Time-Series and Cross-Sectional Data* Kluwer Academic Publishers, Norwell,MA.

Ayers, G. P., Granek, H. & Boers, R. (1997) Ozone in the Marine Boundary Layer at Cape Grim: Model Simulation. *Journal of Atmospheric Chemistry,* 27**,** 179-195.

Ayers, G. P., Penkett, S. A., Gillett, R. W., Bandy, B., Galbally, I. E., Meyer, C. P., Elsworth, C. M., Bentley, S. T. & Forgan, B. W. (1992) Evidence for photochemical control of ozone concentrations in unpolluted marine air. *Nature,* 360**,** 446-449.

Bails, D. G. & Peppers, L. C. (1982) *Business fluctuations: Forecasting techniques and applications,* Englewood Cliffs, NJ, Prentice-Hall.

Ballester, F., Tenías, J. M. & Pérez-Hoyos, S. (2001) Air pollution and emergency hospital admissions for cardiovascular diseases in Valencia, Spain. *Journal of Epidemiology and Community Health,* 55**,** 57-65.

Barnett, A. G., Williams, G. M., Schwartz, J., Neller, A. H., Best, T. L., Petroeschevsky, A. L. & Simpson, R. W. (2005) Air pollution and child respiratory health: a case-crossover study in Australia and New Zealand. *Am J Respir Crit Care Med,* 171**,** 1272-8.

Barnett, V. (2004) *Environmental Statistics: Method and Applications,* Chichester, John Wiley & Sons.

Basagaña, X., Rivera, M., Aguilera, I., Agis, D., Bouso, L., Elosua, R., Foraster, M., de Nazelle, A., Nieuwenhuijsen, M., Vila, J. & Künzli, N. (2012) Effect of the number of measurement sites on land use regression models in estimating local air pollution. *Atmospheric Environment,* 54**,** 634-642.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K. & Briggs, D. J. (2009) Mapping of background air pollution at a fine spatial scale across the European Union. *Science of The Total Environment,* 407**,** 1852-1867.

Beelen, R., Hoek, G., van den Brandt, P. A., Goldbohm, R. A., Fischer, P., Schouten, L. J., Jerrett, M., Hughes, E., Armstrong, B. & Brunekreef, B. (2008) Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR study). *Environ Health Perspect,* 116**,** 196-202.

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q. B., Liu, H. G. Y., Mickley, L. J. & Schultz, M. G. (2001) Global modeling of tropospheric chemistry with

assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research-Atmospheres,* 106**,** 23073-23095.

Bloomfield, P., Royle, J. A., Steinberg, L. J. & Yang, Q. (1996) Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmospheric Environment,* 30**,** 3067-3077.

Böhm, M., McCune, B. & Vandetta, T. (1991) Diurnal curves of tropospheric ozone in the Western United States. *Atmospheric Environment. Part A. General Topics,* 25**,** 1577-1590.

Briggs, D. (2005a) The role of GIS: coping with space (and time) in air pollution exposure assessment. *Journal of toxicology and environmental health. Part A,* 68**,** 1243-1261.

Briggs, D. J. (2005b) Environmental measurment and modelling: geographical information. IN NIEUWENHUIJSEN, M. J. (Ed.) *Exposure assessment in occupational and environmental epidemiology.* New York, Oxford unversity press.

Briggs, D. J. (2007) The use of GIS to evaluate traffic-related pollution. *Occup Environ Med,* 64**,** 1-2.

Briggs, D. J. (2008) A framework for integrated environmental health impact assessment of systemic risks. *Environ Health,* 7**,** 61.

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K. & Van der Veen, A. (1997) Mapping urban air pollution using: a rigression-based approch. *International Journal of Geographical Information Science,* 11**,** 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. *Sci Total Environ,* 253**,** 151-67.

Brook, R., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., Smith, S., Tager, I., Population, E. P. o. & Association, P. S. o. t. A. H. (2002a) Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation.,* 109**,** 2655 - 2671.

Brook, R. D., Brook, J. R., Urch, B., Vincent, R., Rajagopalan, S. & Silverman, F. (2002b) Inhalation of fine particulate air pollution and ozone causes acute arterial vasoconstriction in healthy adults. *Circulation,* 105**,** 1534-6.

Brunekreef, B., Smit, J., De Jongste, J., Neijens, H., Gerritsen, J., Postma, D., Aalberse, R., Koopman, L., Kerkhof, M., Wijga, A. & Van Strien, R. (2002) The Prevention and Incidence of Asthma and Mite Allergy (PIAMA) birth cohort study: Design and first results. *Pediatric Allergy and Immunology,* 13**,** 55-60.

Burrough, P. A. & McDonnel, R. A. (1998) Creating continuous surface from point data. IN BURROUGH, P. A. & MCDONNEL, R. A. (Eds.) *Principles of Geographical Information System.* Oxford, Oxford University Press .

Byun, D. & Schere, K. L. (2004) Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. *Applied Mechanics Reviews* 59**,** 51.

Caballero, S., Galindo, N., Pastor, C., Varea, M. & Crespo, J. (2007) Estimated tropospheric ozone levels on the southeast Spanish Mediterranean coast. *Atmospheric Environment,* 41**,** 2881-2886.

Carter, W. P. L. (1991) Development of ozone reactivity scales for volatile organic compounds. *Other Information: Sponsored by Environmental Protection Agency, Research Triangle Park, NC. Atmospheric Research and Exposure Assessment Lab.*

Cattell, R. B. (1966) The Scree Test For The Number Of Factors. *Multivariate Behavioral Research,* 1**,** 245-276.

Chameides, W., Lindsay, R., Richardson, J. & Kiang, C. (1988) The role of biogenic hydrocarbons in urban photochemical smog: Atlanta as a case study. *Science,* 241**,** 1473-1475.

Chance, K. V., Burrows, J. P., Perner, D. & Schneider, W. (1997) Satellite measurements of atmospheric ozone profiles, including tropospheric ozone, from ultraviolet/visible

measurements in the nadir geometry: a potential method to retrieve tropospheric ozone. *Journal of Quantitative Spectroscopy and Radiative Transfer,* 57**,** 467-476.

Chang, J. C. & Hanna, S. R. (2004) Air quality model performance evaluation. *Meteorology and Atmospheric Physics,* 87**,** 167-196.

Chemel, C., Sokhi, R. S., Yu, Y., Hayman, G. D., Vincent, K. J., Dore, A. J., Tang, Y. S., Prain, H. D. & Fisher, B. E. A. (2010) Evaluation of a CMAQ simulation at high resolution over the UK for the calendar year 2003. *Atmospheric Environment,* 44**,** 2927-2939.

Cleveland, W. S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association,* 74**,** 829-836.

Cleveland, W. S. & Devlin, S. J. (1988) Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association,* 83**,** 596-610.

Collins, S. (1998) Modeling urban spatial variation in air quality using GIS. . IN GATRELL, A. & LOYTONEN, M. (Eds.) *GIS and health. GIS Data.* London, Taylor and Francis.

Colvile, R. & Briggs, D. (2000) Dispersion modelling. *Spatial epidemiology: methods and applications.* Oxford, Oxford University Press.

Cotgreave, I. (1996) Absorption and metabolic fate of ozone the molecular basis of ozone -induced toxicity. *Scandinavian Journal of Work, Environment and Health* 22(suppl.3) 15-26.

Coyle, M., Smith, R. I., Stedman, J. R., Weston, K. J. & Fowler, D. (2002) Quantifying the spatial distribution of surface ozone concentration in the UK. *Atmospheric Environment,* 36**,** 1013-1024.

Cyrys, J., Hochadel, M., Gehring, U., Hoek, G., Diegmann, V., Brunekreef, B. & Heinrich, J. (2005) GIS-Based Estimation of Exposure to Particulate Matter and NO2 in an Urban Area: Stochastic versus Dispersion Modeling. National Institute of Environmental Health Sciences.

Dabdub, D., DeHaan, L. L. & Seinfeld, J. H. (1999) Analysis of ozone in the San Joaquin Valley of California. *Atmospheric Environment,* 33**,** 2501-2514.

Dadvand, P., Rankin, J., Rushton, S. & Pless-Mulloli, T. (2011) Ambient air pollution and congenital heart disease: A register-based study. *Environmental Research,* 111**,** 435-441.

Damsleth, E. & Spjotvoll, E. (1982) Estimation of Trigonometric Components in Time-Series. *Journal of the American Statistical Association,* 77**,** 381-387.

Daniel, Q. T. & Denise, L. M. (2006) Spatial variability of summertime tropospheric ozone over the continental United States: Implications of an evaluation of the CMAQ model. *Atmospheric Environment,* 40**,** 3041-3056.

Davies, T. D., Kelly, P. M., Low, P. S. & Pierce, C. E. (1992) Surface Ozone Concentrations in Europe: Links With the Regional-Scale Atmospheric Circulation. *J. Geophys. Res.,* 97**,** 9819-9832.

Davison, S., Elshout, S. & Wester, B. (2011) Integrated Urban Emission Inventories. IN AIR, C. I. T. E. (Ed.) *Citeair II:.* EUROPEAN UNION.

de Hoogh, K. (1999) Exposure to traffic-related air pollution within a GIS environment. University College Narthampton and University of Leicester

Debella-Gilo, M. & Etzelmüller, B. (2009) Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. *CATENA,* 77**,** 8-18.

Derwent, R. G. (1995) Sources, Distributions, and Fates of VOCs in the Atmosphere. *Environmental Science and Technology,* 4**,** 1-15.

Derwent, R. G. (2001) Transient Behaviour of Tropospheric Ozone Precursors in a Global 3-D CTM and Their Indirect Greenhouse Effects. *Climatic Change,* 49**,** 463-487.

Derwent, R. G., Grennfelt, P., Hov, O., Langner, J., Lindskog, A. & Solberg, S. (2005) The development of European surface ozone . Implications for a revised abtement policy: A contribution from the EU research project NEPAP. Norway, Nowegian Institute for Air Research.

Derwent, R. G., Stevenson, D. S., Doherty, R. M., Collins, W. J. & Sanderson, M. G. (2008) How is surface ozone in Europe linked to Asian and North American NOx emissions? *Atmospheric Environment,* 42**,** 7412-7422.

Diem, J. E. (2003) A critical examination of ozone mapping from a spatial-scale perspective. *Environmental Pollution,* 125**,** 369-383.

Diem, J. E. & Comrie, A. C. (2001) Allocating anthropogenic pollutant emissions over space: application to ozone pollution management. *Journal of Environmental Management,* 63**,** 425-447.

Diem, J. E. & Comrie, A. C. (2002) Predictive mapping of air pollution involving sparse spatial observations. *Environmental Pollution,* 119**,** 99-117.

Dilks, D. W., Canale, R. P. & Meier, P. G. (1992) Development of Bayesian Monte Carlo techniques for water quality model uncertainty. *Ecological Modelling,* 62**,** 149-162.

DOU, Y., Nhu D, L. & James V, Z. (2010) MODELING HOURLY OZONE CONCENTRATION FIELDS. *A Journal Devoted To All Areas Of Applied Statistics,* 4**,** 1183-1213.

Dubus, I. G., Brown, C. D. & Beulke, S. (2003) Sources of uncertainty in pesticide fate modelling. *Science of The Total Environment,* 317**,** 53-72.

Duenas, C., Fernandez, M. C., Canete, S., Carretero, J. & Liger, E. (2004) Analyses of ozone in urban and rural sites in Malaga (Spain). *Chemosphere,* 56**,** 631-9.

Dueñas, C., Fernández, M. C., Cañete, S., Carretero, J. & Liger, E. (2002) Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. *Science of The Total Environment,* 299**,** 97-113.

Dugandzic, R., Dodds, L., Stieb, D. & Smith-Doiron, M. (2006) The association between low level exposures to ambient air pollution and term low birth weight: a retrospective cohort study. *Environmental Health: A Global Access Science Source,* 5**,** 3.

Earnest, A., Chen, M., Ng, D. & Sin, L. (2005) Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research,* 5**,** 36.

EEA (2001) Air pollution by ozonein Europe in summer 2001:Overview of exceedances of EC ozone thresholdvalues during the summer season April–August 2001. IN REPORT, T. (Ed.) *13/2001.* Copenhagen K, Denmark, European Env. Agency.

EEA (2008) Impact of Europe's changing climate 2008 indicator-based assessment. *EEA Report.* Copenhagen.

EEA (2009) Air pollution by ozone across Europe during 2008:overview of exceedance of EC ozone threshold values for April-September 2008. IN 2/2009 (Ed.) *Technical report.* Copenhagen k, Denmark, European Env. Agency.

EEA (2011) Air quality in Europe. *Technical report.* Copenhagen, Denmark, European Environment Agency.

EEA (2012) Air pollution by ozone across Europe during summer 2011: overview of exceedances of EC ozone threshold values for April-September 2011. IN NO1 (Ed.) *Technical report.* Copenhagen, Denmark, European Environment Agency.

El Raey, M., Shalaby, E. A., Ghatass, Z. F. & Marey, H. S. (2006) Time series analysis of ambient air concentrations in Alexandria and Nile delta region, Egypt. *The 2nd Environmental Physical Conference.* Alexandria, Egypt.

Elampari, K. & Chithambarathanu, T. (2011) Diurnal and Seasonal Variations in Surface Ozone Levels at Tropical Semi- Urban site ,Nagercoil , India, and Relationships with Meteorological Conditions. *International Journal of Science and Technology,* 1**,** 80-88.

Fernández-Fernández, M. I., Gallego, M. C., García, J. A. & Acero, F. J. (2011) A study of surface ozone variability over the Iberian Peninsula during the last fifty years. *Atmospheric Environment,* 45**,** 1946-1959.

Field, A. (2009) *Discovering statistics using SPSS,* London, SAGA Publications Ltd

Finlayson-Pitts, B. J. & Pitts Jr, J. N. (1986) *Atmospheric Chemistry: Fundamental and Experimental Techniques,* New york, John Wiley.

Finlayson-Pitts, B. J. & Pitts Jr, J. N. (2000a) Chapter 1 - Overview of the Chemistry of Polluted and Remote Atmospheres. *Chemistry of the Upper and Lower Atmosphere.* San Diego, Academic Press.

Finlayson-Pitts, B. J. & Pitts Jr, J. N. (2000b) Chapter 2 - The Atmospheric System. *Chemistry of the Upper and Lower Atmosphere.* San Diego, Academic Press.

Fischer, P. H., Brunekreef, B. & Lebret, E. (2004) Air pollution related deaths during the 2003 heat wave in the Netherlands. *Atmospheric Environment,* 38**,** 1083-1085.

Fishman, J. & Balok, A. E. (1999) Calculation of daily tropospheric ozone residuals using TOMS and empirically improved SBUV measurements: Application to an ozone pollution episode over the eastern United States. *J. Geophys. Res.,* 104**,** 30319-30340.

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,* 39**,** 6121-6129.

Fowler, D., Flechard, C., Skiba, U., Coyle, M. & Cape, J. N. (1998) The Atmospheric Budget of Oxidized Nitrogen and Its Role in Ozone Formation and Deposition. *New Phytologist,* 139**,** 11-23.

Fox, D. G. (1981) Judging Air Quality Model Performance. *Bulletin of the American Meteorological Society,* 62**,** 599-609.

Fusco, D., Forastiere, F., Michelozzi, P., Spadea, T., Ostro, B., Arca, M. & Perucci, C. A. (2001) Air pollution and hospital admissions for respiratory conditions in Rome, Italy. *Eur Respir J,* 17**,** 1143-50.

Galizia, A. & Kinney, P. L. (1999) Long-term residence in areas of high ozone: associations with respiratory health in a nationwide sample of nonsmoking young adults. *Environ Health Perspect,* 107.

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.

Gariazzo, C., Silibello, C., Finardi, S., Radice, P., Piersanti, A., Calori, G., Cecinato, A., Perrino, C., Nussio, F., Cagnoli, M., Pelliccioni, A., Gobbi, G. P. & Di Filippo, P. (2007) A gas/aerosol air pollutants study over the urban area of Rome using a comprehensive chemical transport model. *Atmospheric Environment,* 41**,** 7286-7303.

Gehring, U., van Eijsden, M., Dijkema, M. B. A., van der Wal, M. F., Fischer, P. & Brunekreef, B. (2011a) Traffic-related air pollution and pregnancy outcomes in the Dutch ABCD birth cohort study. *Occupational and Environmental Medicine,* 68**,** 36-43.

Gehring, U., Wijga, A. H., Fischer, P., de Jongste, J. C., Kerkhof, M., Koppelman, G. H., Smit, H. A. & Brunekreef, B. (2011b) Traffic-related air pollution, preterm birth and term birth weight in the PIAMA birth cohort study. *Environmental Research,* 111**,** 125-135.

Gilliland, F., Avol, E., Kinney, P., Jerrett, M., Dvonch, T., Lurmann, F., Buckley, T., Breysse, P., Keeler, G., de Villiers, T. & McConnell, R. (2005) Air Pollution Exposure Assessment for Epidemiologic Studies of Pregnant Women and Children: Lessons Learned from the Centers for Children's Environmental Health and Disease Prevention Research. *Environmental Health Perspectives,* 113**,** 1447.

Griffith, D. A. & Layne., L. J. (1999) *A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Data Sets,* New york, Oxford University Press.

Guenther, A., Hewitt, C. N., Erickson, D., Fall, R., Geron, C., Graedel, T., Harley, P., Klinger, L., Lerdau, M., McKay, W. A., Pierce, T., Scholes, B., Steinbrecher, R., Tallamraju, R., Taylor, J. & Zimmerman, P. (1995) A global model of natural volatile organic compound emissions. *J. Geophys. Res.,* 100**,** 8873-8892.

Gulliver, J., de Hoogh, K., Fecht, D., Vienneau, D. & Briggs, D. (2011) Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. *Atmospheric Environment,* 45**,** 7072-7080.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C., , (1998) *Multivariate Data*

Analysis, Prentice-Hall International, Inc.

Hanna, S. & Chang, J. (2012) Acceptance criteria for urban dispersion model evaluation. *Meteorology and Atmospheric Physics,* 116**,** 133-146.

Hansen, C., Neller, A., Williams, G. & Simpson, R. (2007) Low levels of ambient air pollution during pregnancy and fetal growth among term neonates in Brisbane, Australia. *Environmental Research,* 103**,** 383-389.

Hansen, C. A., Barnett, A. G., Jalaludin, B. B. & Morgan, G. G. (2009) Ambient Air Pollution and Birth Defects in Brisbane, Australia. *PLoS ONE,* 4**,** e5408.

Harlap, S. (1974) A TIME-SERIES ANALYSIS OF THE INCIDENCE OF DOWN'S SYNDROME IN WEST JERUSALEM. *American Journal of Epidemiology,* 99**,** 210-217.

Hathout, E. H., Beeson, W. L., Ischander, M., Rao, R. & Mace, J. W. (2006) Air pollution and type 1 diabetes in children. *Pediatric Diabetes,* 7**,** 81-87.

Hayman, J. D., Jenkin, M. E., Pilling, M. J. & Derwent, R. G. (2002) Modelling of Tropspheric Ozone Formation. Department for Environment, Food and Rural Affairs.

Hazucha, M. J. & Lefohn, A. S. (2007) Nonlinearity in human health response to ozone: Experimental laboratory considerations. *Atmospheric Environment,* 41**,** 4559-4570.

Henderson, S. B., Beckerman, B., Jerrett, M. & Brauer, M. (2007) Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. *Environmental Science & Technology,* 41**,** 2422-2428.

Henne, S., Brunner, D., Folini, D., Solberg, S., Klausen, J. & Buchmann, B. (2010) Assessment of parameters describing representativeness of air quality in-situ measurement sites. Copernicus GmbH.

Hennekens, C. H. (1998) Increasing Burden of Cardiovascular Disease : Current Knowledge and Future Directions for Research on Risk Factors. *Circulation,* 97**,** 1095-1102.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P. & Briggs, D. (2008) A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment,* 42**,** 7561-7578.

Hoek, G., Beelen, R., Kos, G., Dijkema, M., van der Zee, S. C., Fischer, P. H. & Brunekreef, B. (2011) Land use regression model for ultrafine particles in Amsterdam. *Environmental science & technology,* 45**,** 622-8.

Hoek, G., Brunekreef, B., Fischer, P. & Wijnen, J. v. (2001) The Association between Air Pollution and Heart Failure, Arrhythmia, Embolism, Thrombosis, and Other Cardiovascular Causes of Death in a Time Series Study. *Epidemiology,* 12**,** 355-357.

Holmes, N. S. & Morawska, L. (2006) A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmospheric Environment,* 40**,** 5902-5928.

Hsu, S. A. (1988) *Coastal meteorology,* San Diego, USA, Academic Press Inc.

Hubbard, M. C. & Cobourn, W. G. (1998) Development of a regression model to forecast ground-level ozone concentration in Louisville, KY. *Atmospheric Environment,* 32**,** 2637-2647.

Ihorst, G., Frischer, T., Horak, F., Schumacher, M., Kopp, M., Forster, J., Mattes, J. & Kuehr, J. (2004) Long- and medium-term ozone effects on lung growth including a broad spectrum of exposure. *Eur Respir J,* 23**,** 292-299.

Jakubauskas, M. E., Legates, D. R. & Kastens, J. H. (2002) Crop identification using harmonic analysis of time-series AVHRR NDVI data. *Computers and Electronics in Agriculture,* 37**,** 127-139.

Jenkin, M. E., Davies, T. J. & Stedman, J. R. (2002) The origin and day-of-week dependence of photochemical ozone episodes in the UK. *Atmospheric Environment,* 36**,** 999-1012.

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. *J Expo Anal Environ Epidemiol,* 15**,** 185-204.

Jerrett, M., Burnett, R. T., Kanaroglou, P., Eyles, J., Finkelstein, N., Giovis, C. & Brook, J. (2007) Modeling the intraurban variability of ambient traffic pollution in Toronto,Canada. *Journal of Toxicology and Environmental Health, Part A,* 70**,** 200-212.

Jerrett, M., Burnett, R. T., Pope, C. A., 3rd, Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E. & Thun, M. (2009) Long-term ozone exposure and mortality. *N Engl J Med,* 360**,** 1085-95.

Jolliffe, I. T. (2002) *Principal Component Analysis (2nd ed.),.* New York, Springer-Verlag.

Joly, M. & Peuch, V.-H. (2012) Objective classification of air quality monitoring sites over Europe. *Atmospheric Environment,* 47**,** 111-123.

Jonson , J., Simpson , D., Fagerli , H. & Solberg , S. (2006) Can weexplain the trends in European ozone levels? *Atmospheric Chemistry and physics ,,* 6.

Kaiser, H. F. (1960) The application of electronic computers to factor analysis. *Educational and Psychlogical Measurment,* 20**,** 141-151.

Karakatsani, A., Kapitsimadis, F., Pipikou, M., Chalbot, M. C., Kavouras, I. G., Orphanidou, D., Papiris, S. & Katsouyanni, K. (2010) Ambient air pollution and respiratory health effects in mail carriers. *Environmental Research,* 110**,** 278-285.

Kassomenos, P., Papaloukas, C., Petrakis, M. & Karakitsios, S. (2008) Assessment and prediction of short term hospital admissions: the case of Athens, Greece. *Atmospheric Environment,* 42**,** 7078-7086.

Kempen, B., Brus, D. J., Heuvelink, G. B. M. & Stoorvogel, J. J. (2009) Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma,* 151**,** 311-326.

Kim, S.-W., Yoon, S.-C., Won, J.-G. & Choi, S.-C. (2007) Ground-based remote sensing measurements of aerosol and ozone in an urban area: A case study of mixing height evolution and its effect on ground-level ozone concentrations. *Atmospheric Environment,* 41**,** 7069-7081.

Klingberg, J., Karlsson, P. E., Pihl Karlsson, G., Hu, Y., Chen, D. & Pleijel, H. (2012) Variation in ozone exposure in the landscape of southern Sweden with consideration of topography and coastal climate. *Atmospheric Environment,* 47**,** 252-260.

Kumar, U. & Jain, V. K. (2010) ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). *Stochastic Environmental Research and Risk Assessment,* 24**,** 751-760.

Kyriakidis, P. C. & Journel, A. G. (1999) Geostatistical Space-Time Models: A Review. *Mathematical Geology,* 31**,** 651-684.

Larssen, S., Sluyter, R. & Helimis, C. (1999) Criteria for EUROAIRNET-The EEA air quality monitoring and information network. *EuropeanTopic Centre for Air Quality ETC-AP.* Copenhagen.

Le, H. Q., Batterman, S. A., Wirth, J. J., Wahl, R. L., Hoggatt, K. J., Sadeghnejad, A., Hultin, M. L. & Depa, M. (2012) Air pollutant exposure and preterm and term small-for-gestational-age births in Detroit, Michigan: Long-term trends and associations. *Environment International,* 44**,** 7-17.

Lee, S. J., Hajat, S., Steer, P. J. & Filippi, V. (2008) A time-series analysis of any short-term effects of meteorological and air pollution factors on preterm births in London, UK. *Environmental Research,* 106**,** 185-194.

Leech, N. L., Barrett, K. C. & Morgan, G. A. (2008) *SPSS for Intermediate Statistics: Use and Interpolation,* New York, Tylor & Francis Group, LLC.

Lelieveld, J. & Crutzen, P. J. (1991) The role of clouds in tropospheric photochemistry. *Journal of Atmospheric Chemistry,* 12**,** 229-267.

Levy, H., II, Mahlman, J. D., Moxim, W. J. & Liu, S. C. (1985) Tropospheric Ozone: The Role of Transport. *J. Geophys. Res.,* 90**,** 3753-3772.

Logan, J. A. (1985) Tropospheric Ozone: Seasonal Behavior, Trends, and Anthropogenic Influence. *J. Geophys. Res.,* 90**,** 10463-10482.

Long, X., Pauws, S., Pijl, M., Lacroix, J., Goris, A. & Aarts, R. (2009) Analysis and prediction of daily physical activity level data using autoregressive integrated moving average models. S.l., s.n.

Lou Thompson, M., Reynolds, J., Cox, L. H., Guttorp, P. & Sampson, P. D. (2001) A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment,* 35**,** 617-630.

Madsen, C., Carlsen, K. C. L., Hoek, G., Oftedal, B., Nafstad, P., Meliefste, K., Jacobsen, R., Nystad, W., Carlsen, K.-H. & Brunekreef, B. (2007) Modeling the intra-urban variability of outdoor traffic pollution in Oslo, Norway—A GA2LEN project. *Atmospheric Environment,* 41**,** 7500-7511.

Makridakis, S. & Hibon, M. (1997) ARMA Models and the Box–Jenkins Methodology. *Journal of Forecasting,* 16**,** 147-163.

Marr, L. C. & Harley, R. A. (2002) Spectral analysis of weekday–weekend differences in ambient ozone, nitrogen oxide, and non-methane hydrocarbon time series in California. *Atmospheric Environment,* 36**,** 2327-2335.

Martin, R. V., Jacob, D. J., Logan, J. A., Bey, I., Yantosca, R. M., Staudt, A. C., Li, Q., Fiore, A. M., Duncan, B. N., Liu, H., Ginoux, P. & Thouret, V. (2002) Interpretation of TOMS observations of tropical tropospheric ozone with a global model and in situ observations. *J. Geophys. Res.,* 107**,** 4351.

Massman, W. J. & Grantz , D. A. (1995) Estimating canopy conductance to ozone uptake from observations of evapotranspiration at the canopy scale and at the leaf scale*. *Global Change Biology,* 1**,** 183-198.

Maucha, G. & Büttner, G. (2005) Validation of the European CORINE Land Cover 2000 database. *Global Developments in Environmental Earth Observation from Space***,** 449-457.

Mayer, H. (1999) Air pollution in cities. *Atmospheric Environment,* 33**,** 4029-4037.

McDonnell, W. F., Abbey, D. E., Nishino, N. & Lebowitz, M. D. (1999) Long-Term Ambient Ozone Concentration and the Incidence of Asthma in Nonsmoking Adults: The Ahsmog Study. *Environmental Research,* 80**,** 110-121.

McGregor, G. R. (1996) Identification of air quality affinity areas in Birmingham, UK. *Applied Geography,* 16**,** 109-122.

Mcgregor, G. R. & Bamzelis, D. (1995) Synoptic Typing and Its Application to the Investigation of Weather Air-Pollution Relationships, Birmingham, United-Kingdom. *Theoretical and Applied Climatology,* 51**,** 223-236.

Medina, S., Le Tertre, A., Quénel, P., Le Moullec, Y., Lameloise, P., Guzzo, J. C., Festy, B., Ferry, R. & Dab, W. (1997) Air Pollution and Doctors' House Calls: Results from the ERPURS System for Monitoring the Effects of Air Pollution on Public Health in Greater Paris, France, 1991–1995. *Environmental Research,* 75**,** 73-84.

Meleux, F., Solmon, F. & Giorgi, F. (2007) Increase in summer European ozone amounts due to climate change. *Atmospheric Environment,* 41**,** 7577-7587.

Millán, M. M., Mantilla, E., Salvador, R., Carratalá, A., Sanz, M. J., Alonso, L., Gangoiti, G. & Navazo, M. (2000) Ozone Cycles in the Western Mediterranean Basin: Interpretation of Monitoring Data in Complex Coastal Terrain. *Journal of Applied Meteorology,* 39**,** 487-508.

Mills, C. A. (1957) Respiratory and Cardiac Deaths in Los Angeles Smogs. *The American Journal of the Medical Sciences,* 233**,** 379-386.

Ministry of Health (1954) Mortality and morbidity during the London fog of December 1952. *Reports on public health and medical subjects.* London, UK.

Monks, P. S. (2000) A review of the observations and origins of the spring ozone maximum. *Atmospheric Environment,* 34**,** 3545-3561.

Monks, P. S., Granier, C., Fuzzi, S., Stohl, A., Williams, M. L., Akimoto, H., Amann, M., Baklanov, A., Baltensperger, U., Bey, I., Blake, N., Blake, R. S., Carslaw, K., Cooper, O. R., Dentener, F., Fowler, D., Fragkou, E., Frost, G. J., Generoso, S., Ginoux, P., Grewe, V., Guenther, A., Hansson, H. C., Henne, S., Hjorth, J., Hofzumahaus, A., Huntrieser, H., Isaksen, I. S. A., Jenkin, M. E., Kaiser, J., Kanakidou, M., Klimont, Z., Kulmala, M., Laj, P., Lawrence, M. G., Lee, J. D., Liousse, C., Maione, M., McFiggans, G., Metzger, A., Mieville, A., Moussiopoulos, N., Orlando, J. J., O'Dowd, C. D., Palmer, P. I., Parrish, D. D., Petzold, A., Platt, U., Pöschl, U., Prévôt, A. S. H., Reeves, C. E., Reimann, S., Rudich, Y., Sellegri, K., Steinbrecher, R., Simpson, D., ten Brink, H., Theloke, J., van der Werf, G. R., Vautard, R., Vestreng, V., Vlachokostas, C. & von Glasow, R. (2009) Atmospheric composition change – global and regional air quality. *Atmospheric Environment,* 43**,** 5268-5350.

Moral García, F. J., Valiente González, P. & López Rodríguez, F. (2008) Geostatistical Analysis and Mapping of Groundlevel Ozone in a Medium Sized Urban Area. *World Academy of Science, Engineering and Technology*.

Morgan, G., Corbett, S. & Wlodarczyk, J. (1998) Air pollution and hospital admissions in Sydney, Australia, 1990 to 1994. *American Journal of Public Health,* 88**,** 1761-1766.

Morgan, G. A., Vaske, J. J., Gliner, J. A., Harmon, R. J. & Harmon, R. J. (2003) Logistic Regression and Discriminant Analysis: Use and Interpretation. *Journal of the American Academy of Child and Adolescent Psychiatry,* 42**,** 994-997.

Müller, J.-F. (1992) Geographical Distribution and Seasonal Variation of Surface Emissions and Deposition Velocities of Atmospheric Trace Gases. *J. Geophys. Res.,* 97**,** 3787-3804.

Muller, J., Abela, G., Nesto, R. & Tofler, G. (1994) Triggers, acute risk factors and vulnerable plaques: the lexicon of a new frontier. *J Am Coll Cardiol,* 23**,** 809 - 813.

Mustafa, M. G. (1990) Biochemical basis of ozone toxicity. *Journal Name: Free Radical Biology and Medicine Journal,* 9**,** 245-265.

Mustafić, H., Jabre, P., Caussin, C., Murad, M. H., Escolano, S., Tafflet, M., Périer, M.-C., Marijon, E., Vernerey, D., Empana, J.-P. & Jouven, X. (2012) Main Air Pollutants and Myocardial Infarction:A systematic review and meta-analysis. *JAMA: The Journal of the American Medical Association,* 307**,** 713-721.

Narang, I. (2010) Review Series: What goes around, comes around: childhood influences on later lung health?: Long-term follow-up of infants with lung disease of prematurity. *Chronic Respiratory Disease,* 7**,** 259-269.

Nikiforov, S. V., Aggarwal, M., Nadas, A. & Kinney, P. (1998) Methods for spatial interpolation of long-term ozone concentrations. *Journal of Exposure Analysis and Environmental Epidemiology,* 8**,** 465-481.

Nolle, M., Ellul, R., Heinrich, G. & Güsten, H. (2002) A long-term study of background ozone concentrations in the central Mediterranean—diurnal and seasonal variations on the island of Gozo. *Atmospheric Environment,* 36**,** 1391-1402.

Nowak, D. J., Crane, D. E. & Stevens, J. C. (2006) Air pollution removal by urban trees and shrubs in the United States. *Urban Forestry & Urban Greening,* 4**,** 115-123.

Nuckols, J. R., Ward, M. H. & Jarup, L. (2004) Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies. National Institue of Environmental Health Sciences.

Osmond, C. & Barker, D. J. P. (2000) Fetal, Infant, and Childhood Growth Are Predictors of Coronary Heart Disease, Diabetes, and Hypertension in Adult Men and Women. *Environ Health Perspect,* 108**,** 545-553.

Park, S. K., O'Neill, M. S., Vokonas, P. S., Sparrow, D. & Schwartz, J. (2005) Effects of air pollution on heart rate variability: the VA normative aging study. *Environ Health Perspect,* 113**,** 304-9.

Partonen, T., Haukka, J., Nevanlinna, H. & Lönnqvist, J. (2004) Analysis of the seasonal pattern in suicide. *Journal of Affective Disorders,* 81**,** 133-139.

Pereira, G., Cook, A. G., Haggar, F., Bower, C. & Nassar, N. (2012) Locally derived traffic-related air pollution and fetal growth restriction: a retrospective cohort study. *Occupational and Environmental Medicine*.

Perepu, R., Dostal, D., Garcia, C., Kennedy, R. & Sethi, R. (2012) Cardiac dysfunction subsequent to chronic ozone exposure in rats. *Molecular and Cellular Biochemistry,* 360**,** 339-345.

Phillips, D. L., Lee, E. H., Herstrom, A. A., Hogsett, W. E. & Tingey, D. T. (1997) USE OF AUXILIARY DATA FOR SPATIAL INTERPOLATION OF OZONE EXPOSURE IN SOUTHEASTERN FORESTS. *Environmetrics,* 8**,** 43-61.

Piegorsch, W. W. & Bailer, A. J. (2005) Temporal data and autoregressive modelling. *Analyzing Environmental Data.* Chichester, England, John Wiley & Sons,Ltd.

Ponce de Leon, A., Anderson, H. R., Bland, J. M., Strachan, D. P. & Bower, J. (1996) Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. *J Epidemiol Community Health,* 50**,** s63-70.

Pont, V. & Fontan, J. (2001) Comparison between weekend and weekday ozone concentration in large cities in France. *Atmospheric Environment,* 35**,** 1527-1535.

Porta, D. & Fantini, M. (2007) Prospective cohort studies of newborns in Italy to evaluate the role of environmental and genetic characteristics on common childhood disorders. *The Italian J of Pediatrics.* .

Poumadère, M., Mays, C., Le Mer, S. & Blong, R. (2005) The 2003 Heat Wave in France: Dangerous Climate Change Here and Now. *Risk Analysis,* 25**,** 1483-1494.

Proakis, J. G. & Salehi, M. (2002) *Communication Systems Engineering,* New Jersey, Prentice-Hall.

Ramond, A. (2012) Effects of exposure to air pollution on a marker of systemic inflammation in a sample of the GA2LEN cohort. Imperial college London.

Rao, S. T., Ku, J. Y., Berman, S., Zhang, K. & Mao, H. (2003) Summertime Characteristics of the Atmospheric Boundary Layer and Relationships to Ozone Levels over the Eastern United States. *Pure and Applied Geophysics,* 160**,** 21-55.

Rich, D. Q., Mittleman, M. A., Link, M. S., Schwartz, J., Luttmann-Gibson, H., Catalano, P. J., Speizer, F. E., Gold, D. R. & Dockery, D. W. (2006) Increased risk of paroxysmal atrial fibrillation episodes associated with acute increases in ambient air pollution. *Environ Health Perspect,* 114**,** 120-3.

Rich, D. Q., Schwartz, J., Mittleman, M. A., Link, M., Luttmann-Gibson, H., Catalano, P. J., Speizer, F. E. & Dockery, D. W. (2005) Association of Short-term Ambient Air Pollution Concentrations and Ventricular Arrhythmias. *Am. J. Epidemiol.,* 161**,** 1123-1132.

Richards, R. P. & Baker, D. B. (2002) Trends in Water Quality in LEASEQ Rivers and Streams (Northwestern Ohio), 1975–1995. *J. Environ. Qual.,* 31**,** 90-96.

Rodríguez, S. & Guerra, J.-C. (2001) Monitoring of ozone in a marine environment in Tenerife (Canary Islands). *Atmospheric Environment,* 35**,** 1829-1841.

Ross, Z., English, P. B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S. & Jerrett, M. (2005) Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *J Expos Sci Environ Epidemiol,* 16**,** 106-114.

Ruidavets, J., Cournot, M., Cassadou, S., Giroux, M., Meybeck, M. & Ferrieres, J. (2005) Ozone air pollution is associated with acute myocardial infarction. *Circulation,* 111**,** 563 - 569.

Ryan, P. H. (2007) A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environmental Health Perspectives,* 115**,** 278.

Ryerson, T. B., Trainer, M., Holloway, J. S., Parrish, D. D., Huey, L. G., Sueper, D. T., Frost, G. J., Donnelly, S. G., Schauffler, S., Atlas, E. L., Kuster, W. C., Goldan, P. D., Hübler, G., Meagher, J. F. & Fehsenfeld, F. C. (2001) Observations of Ozone Formation in Power Plant Plumes and Implications for Ozone Control Strategies. *Science,* 292**,** 719-723.

Sahsuvaroglu, T., Kanaroglou, P., Finkelstein, N., Newbold, K. B., Jerrett, M., Beckerman, B., Brook, J. R., Finkelstein, M. & Gilbert, D. (2006) A land use regression model for predicting ambient

concentrations of nitrogen dioxide in Hamilton, Ontario, Canada. . *Journal of Air Waste Management Associations,* 56**,** 1059-1069.

Sahu, S. K. (2011) Hierarchical Bayesian models for space-time air pollution data. *Handbook of Statistics: Time Series Analysis, Methods and Applications.* Elsevier.

Salam, M. T., Millstein, J., Li, Y.-F., Lurmann, F. W., Margolis, H. G. & Gilliland, F. D. (2005) Birth outcomes and prenatal exposure to ozone, carbon monoxide, and particulate matter: results from the Children's Health Study. *Environ Health Perspect,* 113**,** 1638 - 1644.

Sanchez, M. L. & Sanz, J. (1994) Application of discriminant analysis to interpret the behaviour of photochemical oxidants in an urban area. *Atmospheric Environment,* 28**,** 1147-1157.

Sargazi, S., Hamid, T. S., Majid, H. & Melika, S. (2011) Application of GIS for the modeling of spatial distribution of air pollutants in Tehran. *Proc. ,* SPIE 8181, 81810I .

Scheel, H. E., Areskoug, H., Geiss, H., Gomiscek, B., Granby, K., Haszpra, L., Klasinc, L., Kley, D., Laurila, T., Lindskog, A., Roemer, M., Schmitt, R., Simmonds, P., Solberg, S. & Toupance, G. (1997) On the Spatial Distribution and Seasonal Variation of Lower-Troposphere Ozone over Europe. *Journal of Atmospheric Chemistry,* 28**,** 11-28.

Schurgers, G., Hickler, T., Miller, P. A. & Arneth, A. (2009) European emissions of isoprene and monoterpenes from the Last Glacial Maximum to present. Copernicus GmbH.

Schwartz, J. (1999) Air Pollution and Hospital Admissions for Heart Disease in Eight U.S. Counties. *Epidemiology,* 10**,** 17-22.

Schwartz, J. (2005) How sensitive is the association between ozone and daily deaths to control for temperature? *Am J Respir Crit Care Med,* 171**,** 627 - 631.

Shan, W., Yin, Y., Zhang, J., Ji, X. & Deng, X. (2009) Surface ozone and meteorological condition in a single year at an urban site in central–eastern China. *Environmental Monitoring and Assessment,* 151**,** 127-141.

Shi, C., Fernando, H. J. S. & Hyde, P. (2012) CMAQ predictions of tropospheric ozone in the U.S. southwest: Influence of lateral boundary and synoptic conditions. *Science of The Total Environment,* 416**,** 374-384.

Sillman, S. (1999) The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments. *Atmospheric Environment,* 33**,** 1821-1845.

Sillman, S. (2003) Tropospheric Ozone and Photochemical Smog. IN EDITORS-IN-CHIEF: HEINRICH, D. H. & KARL, K. T. (Eds.) *Treatise on Geochemistry.* Oxford, Pergamon.

Simpson, D. (1995) Biogenic emissions in Europe 2. Implications for ozone control strategies. *J. Geophys. Res.,* 100**,** 22891-22906.

Simpson, D., Guenther, A., Hewitt, C. N. & Steinbrecher, R. (1995) Biogenic emissions in Europe 1. Estimates and uncertainties. *J. Geophys. Res.,* 100**,** 22875-22890.

Simpson, D., Winiwarter, W., Börjesson, G., Cinderby, S., Ferreiro, A., Guenther, A., Hewitt, C. N., Janson, R., Khalil, M. A. K., Owen, S., Pierce, T. E., Puxbaum, H., Shearer, M., Skiba, U., Steinbrecher, R., Tarrasón, L. & Öquist, M. G. (1999) Inventorying emissions from nature in Europe. *J. Geophys. Res.,* 104**,** 8113-8152.

Singh, V., Carnevale, C., Finzi, G., Pisoni, E. & Volta, M. (2011) A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations. *Environmental Modelling & Software,* 26**,** 778-786.

Skene, K. J., Gent, J. F., McKay, L. A., Belanger, K., Leaderer, B. P. & Holford, T. R. (2010) Modeling effects of traffic and landscape characteristics on ambient nitrogen dioxide levels in Connecticut. *Atmospheric Environment,* 44**,** 5156-5164.

Snel, S. (2004) Improvement of classifications European montoring stations for AIRBASE. A quality control. . European Topic Centre on Air and Climate Change.

Sokhi, R. S., San José, R., Kitwiroon, N., Fragkou, E., Pérez, J. L. & Middleton, D. R. (2006) Prediction of ozone levels in London using the MM5–CMAQ modelling system. *Environmental Modelling & Software,* 21**,** 566-576.

Spangl, W., Schneider, J., Moosmann, L. & Nagl, C. (2007) Draft FInal Report: Representativeness and classification of air quality monitoring stations. IN UMWELTBUNDESAMT (Ed.).

Srebot, V., Gianicolo, E., Rainaldi, G., Trivella, M. & Sicari, R. (2009) Ozone and cardiovascular injury. *Cardiovascular Ultrasound,* 7**,** 30.

Steinbach, V., Ertoz, L. & Kumar, V. (2003) The Challenges of Clustering High Dimensional Data. IN WILLE, L. T. (Ed.) *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition.* New York, Springer-Verlag.

Stevens, J. P. (2002) *Applied Multivariate Statistics for the Social Sciences,* Mahwah,New Jersy, Lawrence Erlbaum Associates,Inc.

Sundberg, J., Karlsson, P.-E., Schenk, L. & Pleijel, H. (2006) Variation in ozone concentration in relation to local climate in south-west Sweden. *Water, Air, & Soil Pollution,* 173**,** 339-354.

Tarasova, O. A. & Karpetchko, A. Y. (2003) Accounting for local meteorological effects in the ozone time-series of Lovozero (Kola Peninsula). HAL - CCSD.

Tiao, G. C. & Grupe, M. R. (1980) Hidden Periodic Autoregressive-Moving Average Models in Time Series Data. *Biometrika,* 67**,** 365-373.

Tong, D. Q. & Mauzerall, D. L. (2006) Spatial variability of summertime tropospheric ozone over the continental United States: Implications of an evaluation of the CMAQ model. *Atmospheric Environment,* 40**,** 3041-3056.

TRS (2008) Ground-level ozone in the 21st century:future trends, impacts and policy implications. IN 15/08 (Ed.) *Science Policy.* London,UK, The Royal Society.

UNECE (2008) Health risks of ozone in Europe. Geneva, United Nations Economic Commission for Europe

Van Eijsden, M., Van Der Wal, M. F. & Bonsel, G. J. (2006) Folic acid knowledge and use in a multi-ethnic pregnancy cohort: the role of language proficiency. *BJOG: An International Journal of Obstetrics & Gynaecology,* 113**,** 1446-1451.

Vandentorren, S., Suzan, F., Medina, S., Pascal, M., Maulpoix, A., Cohen, J. & Ledrans, M. (2004) Mortality in 13 French cities during the August 2003 heat wave. *Am J Public Health,* 94**,** 1518 - 1520.

Vestreng, V., Ntziachristos, L., Semb, A., Reis, S., Isaksen, I. S. A. & TarrasÃ³n, L. (2008) Evolution of NOx emissions in Europe with focus on road transport control measures. *Atmos. Chem. Phys. Discuss.,* 8**,** 10697-10747.

Vienneau, D., de Hoogh, K., Beelen, R., Fischer, P., Hoek, G. & Briggs, D. (2010) Comparison of land-use regression models between Great Britain and the Netherlands. *Atmospheric Environment,* 44**,** 688-696.

Vienneau, D., de Hoogh, K. & Briggs, D. (2009) A GIS-based method for modelling air pollution exposures across Europe. *Science of The Total Environment,* 408**,** 255-266.

Vijayaraghavan, K., Snell, H. E. & Seigneur, C. (2008) Practical Aspects of Using Satellite Data in Air Quality Modeling. *Environmental Science & Technology,* 42**,** 8187-8192.

Vingarzan, R. & Taylor, B. (2003) Trend analysis of ground level ozone in the greater Vancouver/Fraser Valley area of British Columbia. *Atmospheric Environment,* 37**,** 2159-2171.

von Klot, S. (2011) Equivalence of using nested buffers and concentric adjacent rings as predictors in land use regression models. *Atmospheric Environment,* 45**,** 4108-4110.

Wang, M., Beelen, R., Eeftens, M., Meliefste, K., Hoek, G. & Brunekreef, B. (2012) Systematic Evaluation of Land Use Regression Models for NO2. *Environmental Science & Technology,* 46**,** 4481-4489.

Waser, L. T. & Schwarz, M. (2006) Comparison of large-area land cover products with national forest inventories and CORINE land cover in the European Alps. *International Journal of Applied Earth Observation and Geoinformation,* 8**,** 196-207.

Watkiss, P., Pye, S. & Holland, M. (2005) Baseline scenarios for service contract for carrying out cost-benefit analysis of air quality related issues, in particular in the clean air for Europe (CAFE) programme. UK, AEA Technology Environment.

WHO (2006) Air quality Guidelines: Global update 2005. Particulate matter, ozone, nitrogen dioxide and sulfer dioxide. Copenhagen, World Health Organization.

WHO (2008) Health Risks of Ozone from Long-range Transboundary Air Pollution. Copenhagen, Denmark, WHO Regional Office for Europe.

Wilby, R. L. & Tomlinson, O. J. (2000) The 'Sunday Effect' and weekly cycles of winter weather in the UK. *Weather,* 55**,** 214-222.

Wilhelm, M., Qian, L. & Ritz, B. (2009) Outdoor air pollution, family and neighborhood environment, and asthma in LA FANS children. *Health & Place,* 15**,** 25-36.

Wilks, D. S. (1955) *Statistical methods in atmospheric science*, Academic Press,London

Willett, P. (1988) Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management,* 24**,** 577-597.

Willmott, C. J. (1982) Some Comments on the Evaluation of Model Performance. *Bulletin of the American Meteorological Society,* 63**,** 1309-1313.

Wilson, R. C., Fleming, Z. L., Monks, P. S., Clain, G., Henne, S., Konovalov, I. B., Szopa, S. & Menut, L. (2012) Have primary emission reduction measures reduced ozone across Europe? An analysis of European rural background ozone trends 1996-2005. *Atmos. Chem. Phys.,* 12**,** 437-454.

Wong, T. W., Tam, W. S., Yu, T. S. & Wong, A. H. S. (2002) Associations between daily mortalities from respiratory and cardiovascular diseases and air pollution in Hong Kong, China. *Occupational and Environmental Medicine,* 59**,** 30-35.

Yanosky, J. D., Paciorek, C. J., Schwartz, J., Laden, F., Puett, R. & Suh, H. H. (2008) Spatio-temporal modeling of chronic PM10 exposure for the Nurses' Health Study. *Atmospheric Environment,* 42**,** 4047-4062.

Yarnal, B. (1992) *Synoptic Climatology in Environmental Analysis: A Primer,* London, Belhaven Press.

Zanobetti, A. & Schwartz, J. (2008) Mortality displacement in the association of ozone with mortality: an analysis of 48 cities in the United States. *Am J Respir Crit Care Med,* 177**,** 184-9.

Zhang, Z., Wang, F., Costabile, F., Allegrini, I., Liu, F. & Hong, W. (2012) Interpretation of ground-level ozone episodes with atmospheric stability index measurement. *Environmental Science and Pollution Research,* 19**,** 3421-3429.

Ziemke, J. R., Chandra, S., Duncan, B. N., Froidevaux, L., Bhartia, P. K., Levelt, P. F. & Waters, J. W. (2006) Tropospheric ozone determined from Aura OMI and MLS: Evaluation of measurements and comparison with the Global Modeling Initiative's Chemical Transport Model. *J. Geophys. Res.,* 111**,** D19303.

Ziemke, J. R., Chandra, S., Labow, G., Bhartia, P. K., Froidevaux, L. & Witte, J. C. (2011) A global climatology of tropospheric and stratospheric ozone derived from Aura OMI and MLS measurements. *Atmos. Chem. Phys. Discuss.,* 11.

Appendix A

Miscellaneous

Contents

I. Short-term health impact

1) Respiratory diseases

Most of the epidemiological studies of O_3 to date have focused on short-term health effects, and have been either cross-sectional or time series in design. According to the Task Force on Health (UNECE, 2008), the majority of recent epidemiological studies have reported positive and significant associations between short-term exposure to different O_3 concentrations and increased morbidity and mortality from respiratory diseases. Inhaling O_3, in the short term, can cause a variety of health problems, including lung damage, aggravated asthma, and increased susceptibility to respiratory tract illnesses such as pneumonia and bronchitis. The most consistent associations have been seen with impaired pulmonary function, which was found to be correlated with increased medication usage. Kassomenos et al. (2008) undertook a study investigating daily effects of exposure to O_3 and daily hospital admissions due to respiratory and cardiovascular diseases in Athens, Greece (population ca. 1.5 million) between 1992-2000. Admission records were obtained for 7,435 individuals (4,285 males and 3,151 females). This study has shown that a 10μg/m^3 increase in O_3 concentration was associated with a 7.2% increase in the number of daily hospital admissions.

Ponce de Leon et al. (1996) conducted a 4 year study in London (1987 - 1992), and reported an increase in daily hospital admissions due to respiratory diseases associated with increased O_3 concentration of 58μg/m^3 (the 10th - 90th percentile), with a relative risk of 1.05 (95% Cl, 1.02-1.07) for all ages except for children (0-14 years). In contrast, Anderson et al. (2005) found no significant association between ambient O_3 and respiratory illness for any age group (0-14, 15-65 and >65 years). Similar results were obtained in a study of residents in urban Australia and New Zealand by Barnett et al. (2005). No significant effects were detected for any age group, except with respiratory admissions among children (1-4 years), for whom the excess risk was 3.5% (95% Cl: 1.8-5.2) for a 19.23μg/m^3 increase in daily mean O_3 concentration, during warm months only; overall (for the whole year), there was no effect. Fusco et al. (2001) conducted a similar study in Rome (study period 1995-1997). They found that a 23.9μg/m³ increase in concentration (equivalent to the inter quintile range) was associated with an 8.1% (95% Cl: 2.1-7.3) increase in hospital admission due to acute respiratory infection among children (0-14 years) only.

2) Cardiovascular diseases

Air pollution is recognized as a critical and modifiable determinant of cardiovascular diseases in urban populations. Evidence by Hennekens (1998) also suggested that, when people migrate to a new environment, their risk of cardiovascular disease may be affected (i.e. O_3 concentrations are different from one location to another). Srebot et al. (2009) reviewed literature on the short-term

impacts of O_3 exposure on arterial pressure control, vascular tone, autonomic control of serum concentration of inflammatory markers and heart rate.

Several studies have explored the correlation between increased ambient O_3 concentration and cardiovascular disease. In an animal study, Perepu et al. (2012) observed normal adult rats exposed to 0.8ppm of O_3 compared to filtered air for 8 hours/day for 28 to 56 days. They reported that a significant reduction in myocardial function could be observed in the more highly exposed rats due to increased levels of oxidative stress and inflammation.

Some epidemiological studies, however, have found no association between acute exposure to ambient O_3 concentration and hospital admission due to cardiovascular diseases (WHO, 2008, Anderson et al., 2005); significant associations were observed in a few studies only. One of these was a study conducted in Boston, Massachusetts by Park et al. (2005). The association between reduced heart rate variability (HRV), an indicator of poor cardiac autonomic function, and exposures to 4-hr, 24-hr, and 48-hr moving averages of ambient air pollutants in 497 men seen between November 2000 and October 2003 was explored. The results showed that HRV was reduced by 11.5% (95% CI: 0.4-21.3%) per 26$\mu g/m^3$ increment in 4-hr average O_3. The effect was also stronger in susceptible people (i.e. those suffering from ischemic heart disease (IHD) and hypertension).

A randomised, double-blinded crossover chamber study including 25 healthy adults reported that exposures influenced macrovascular diameter and caused brachial artery narrowing after short-term inhalation (2 hour exposure to both 240$\mu g/m^3$ O_3 and 150$\mu g/m^3$ PM) ((Brook et al., 2002b). A similar response is suspected to occur in the coronary diameter: an impact on healthy adults may occur with a reduction as little as 0.1mm (Srebot et al., 2009). Among susceptible individuals, this degree of vasoconstriction may promote cardiac ischemia or trigger instability of susceptible plaques (Muller et al., 1994). Rich et al. (2006) reported on a case-crossover study conducted in Boston, Massachusetts, exploring the association between O_3 and paroxysmal atrial fibrillation episodes (PAF) using 203 patients with implantable cardioverter defibrillators, followed between 1995 and 1999 until 2002. A significant positive association was observed, with an odds ratio of 2.1 (95% CI: 1.22-3.54) per 43.2$\mu g/m^3$ O_3 during the hour before arrhythmia. The associations were very weak, in contrast, for exposures averaged over the previous 24 hours, and no significant risks were associated with other air pollutants. On the other hand, results of short-term effect studies suggest a link with adverse cardiovascular events such as myocardial infarction (Ruidavets et al., 2005), heart failure (Hoek et al., 2001), and life-threatening arrhythmias (Rich et al., 2005) though the evidence remains inconclusive.

3) Mortality

O_3 is not only a risk factor for increased morbidity but is also estimated to be responsible for ca. 3 million premature deaths world-wide each year, according to the World Health Organization (WHO, 2006). It is also estimated that, in the European Union (25 countries), about 21,000 premature deaths occur annually after days with high O_3 levels (WHO, 2008).

Four meta-analyses have been undertaken of the relationship between O_3 and mortality (UNECE, 2008). These suggested significant, independent associations between O_3 exposures and different causes of mortality. Impacts on respiratory mortality are strongest; those on cardiovascular mortality seem to be weaker. These effects are not influenced by other air pollutants, weather factors (e.g. temperature and humidity), season or modelling strategy (WHO, 2006). Even so, this evidence is not considered sufficient to confirm an association with mortality (UNECE, 2008).

One of the recent studies was conducted by Jerrett et al. (2009) in 96 metropolitan areas in the United States, using health data of 448,850 subjects (and including 118,777 deaths) from the American Cancer Society Cancer Prevention Study cohort II. Associations were sought between daily maximum O_3 concentrations, for the period from 1977 to 2000, and mortality from cardiopulmonary and respiratory diseases. The results showed that, when adjusted for $PM_{2.5}$ concentrations, each 20 $\mu g/m^3$ increase in O_3 concentration was associated with a 4% (95% Cl: 1.0-6.7) increase in risk of death from respiratory diseases, primarily due to pneumonia and chronic obstructive pulmonary disease. In contrast, O_3 had no detectable effects on the risk of death from all causes, cardiopulmonary, ischemic heart and cardiovascular diseases when $PM_{2.5}$ concentration was taken into account.

Another study by Zanobetti and Schwartz (2008) explored effects of O_3 on short-term displacement of death date (so-called harvesting) in 48 cities in the USA between 1989 and 2000. The results indicated that an increase in summer-time O_3 concentration of 20 $\mu g/m^3$, as the 8 hour average, is associated with a 0.3% (95% CI: 0.2-0.4) and 0.5% (95% CI: 0.05-0.96) increase in total mortality, for time lags of 0 and 0-3 days respectively. This suggests that risk assessments based on exposure on a single day are likely to underestimate the health impact of O_3.

A quantitative meta-analysis of peer reviewed studies was conducted by Anderson et al. (2005) using databases of time-series studies from several European cities. A strong, statistically significant association between short-term exposure to O_3 and mortality was found. also evaluated 95 communities in the USA, and found a 0.52% (95% CI: 0.27–0.77%) increase in daily death for a $20\mu g/m^3$ increase in O_3 concentration during the last week. A similar result was recorded from a large study which assessed the impact of O_3 exposure on cause-specific and daily total mortality

from 23 European cities . A ca. 10mg/m^3 increase in 1-hr O_3 concentration was related to a 0.45% (95% CI: 0.17-0.52%) increase in the number of deaths due to cardiovascular diseases.

Mortality in children has not fully been assessed; most effects were detected in elderly people and were seen to be strongest in the warm season. The literature to date is thus inconclusive for the short-term impact. By the same token, it can be argued that, if these studies are to be convincing, they need to be based on more accurate and specific measures of exposure, both in terms of their spatial and temporal resolution.

II. Emission sources of O_3 precursors

Base data, reported in the UNECE/EMEP Nomenclature for Reporting (NFR) sector format are aggregated into the following EEA sector codes to obtain a consistent reporting format across all countries and pollutants:

- Energy production and distribution: emissions from public heat and electricity generation, oil refining, production of solid fuels, extraction and distribution of solid fossil fuels and geothermal energy;
- Energy use in industry: emissions from combustion processes used in the manufacturing industry including boilers, gas turbines and stationary engines;
- Industrial processes: emissions derived from non-combustion related processes such as the production of minerals, chemicals and metal production;
- Road transport: light and heavy duty vehicles, passenger cars and motorcycles;
- Non-road transport: railways, domestic shipping, certain aircraft movements, and non-road mobile machinery used in agriculture & forestry;
- Commercial, institutional and households: emissions principally occurring from fuel combustion in the services and household sectors;
- Solvent and product use: non-combustion related emissions mainly in the services and households sectors including activities such as paint application, dry-cleaning and other use of solvents;
- Agriculture: manure management, fertiliser application, field-burning of agricultural wastes
- Waste: incineration, waste-water management.

III. CLC2000 classes

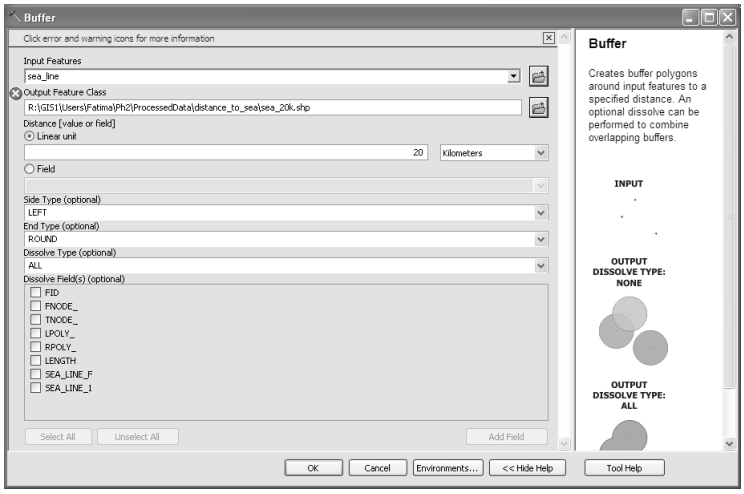
TableA.1 The full descriptive list of CLC2000 classes

GRID_COD	CLC_CODE	LABEL1	LABEL2	LABEL3	new classes
1	111	Artificial surfaces	Urban fabric	Continuous urban fabric	High density residential
2	112	Artificial surfaces	Urban fabric	Discontinuous urban fabric	Low density residential
3	121	Artificial surfaces	Industrial, commercial and transport units	Industrial or commercial units	Industerial
4	122	Artificial surfaces	Industrial, commercial and transport units	Road and rail networks and associated land	Industerial
5	123	Artificial surfaces	Industrial, commercial and transport units	Port areas	Industerial
6	124	Artificial surfaces	Industrial, commercial and transport units	Airports	Industerial
7	131	Artificial surfaces	Mine, dump and construction sites	Mineral extraction sites	Industerial
8	132	Artificial surfaces	Mine, dump and construction sites	Dump sites	Industerial
9	133	Artificial surfaces	Mine, dump and construction sites	Construction sites	Industerial
10	141	Artificial surfaces	Artificial, non-agricultural vegetated areas	Green urban areas	herbaceous
11	142	Artificial surfaces	Artificial, non-agricultural vegetated areas	Sport and leisure facilities	herbaceous
12	211	Agricultural areas	Arable land	Non-irrigated arable land	Agriculture
13	212	Agricultural areas	Arable land	Permanently irrigated land	Agriculture
14	213	Agricultural areas	Arable land	Rice fields	Agriculture
15	221	Agricultural areas	Permanent crops	Vineyards	Agriculture
16	222	Agricultural areas	Permanent crops	Fruit trees and berry plantations	Agriculture
17	223	Agricultural areas	Permanent crops	Olive groves	Agriculture
18	231	Agricultural areas	Pastures	Pastures	herbaceous
19	241	Agricultural areas	Heterogeneous agricultural areas	Annual crops associated with permanent crops	Agriculture
20	242	Agricultural areas	Heterogeneous agricultural areas	Complex cultivation patterns	Agriculture
21	243	Agricultural areas	Heterogeneous agricultural areas	Land principally occupied by agriculture, with significant areas of natural vegetation	Agriculture
22	244	Agricultural areas	Heterogeneous agricultural areas	Agro-forestry areas	Agriculture
23	311	Forest and semi natural areas	Forests	Broad-leaved forest	Forest
24	312	Forest and semi natural areas	Forests	Coniferous forest	Forest
25	313	Forest and semi natural areas	Forests	Mixed forest	Forest
26	321	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Natural grasslands	herbaceous
27	322	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Moors and heathland	herbaceous
28	323	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Sclerophyllous vegetation	herbaceous
29	324	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Transitional woodland-shrub	herbaceous
30	331	Forest and semi natural areas	Open spaces with little or no vegetation	Beaches, dunes, sands	open space
31	332	Forest and semi natural areas	Open spaces with little or no vegetation	Bare rocks	open space
32	333	Forest and semi natural areas	Open spaces with little or no vegetation	Sparsely vegetated areas	open space
33	334	Forest and semi natural areas	Open spaces with little or no vegetation	Burnt areas	open space
34	335	Forest and semi natural areas	Open spaces with little or no vegetation	Glaciers and perpetual snow	open space

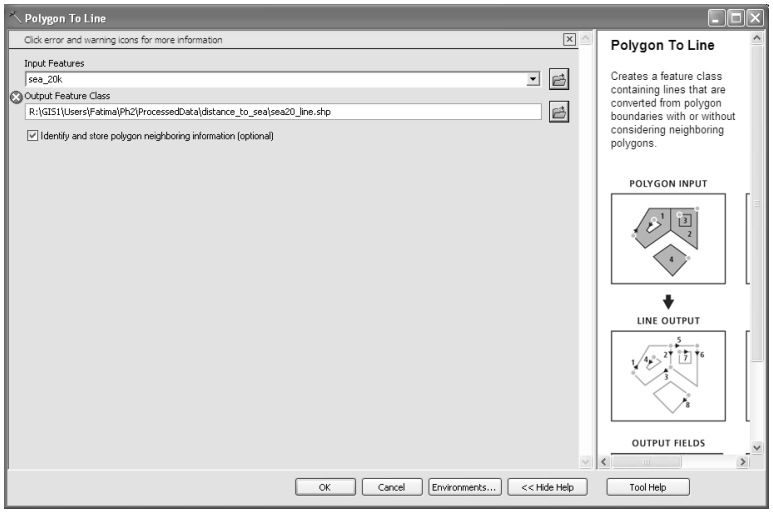
IV. Calculate the distance to sea

First in ARC/map:

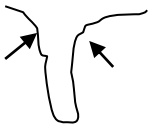
A. Prepare the coast line:
   1- Add corine 523 attribute, converted from polygon to Line (sea_line.shp).
   2- Buffer 20km around the sea_line save it as sea_20k.shp



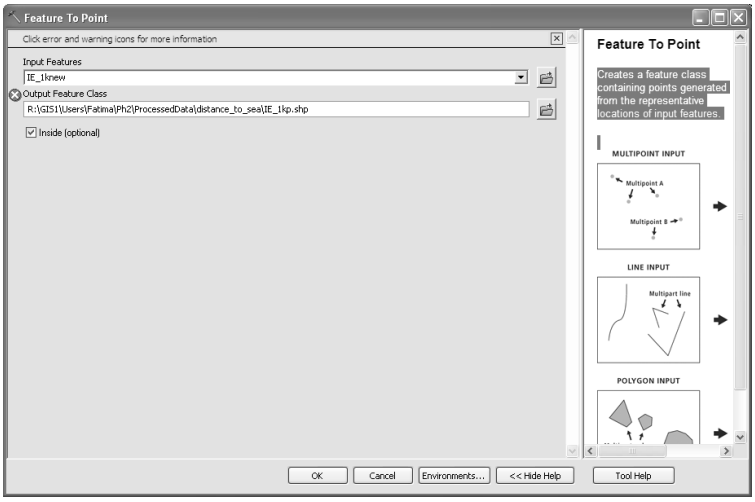
   3- Convert the polygon sea_20k.shp to line sea20_line by



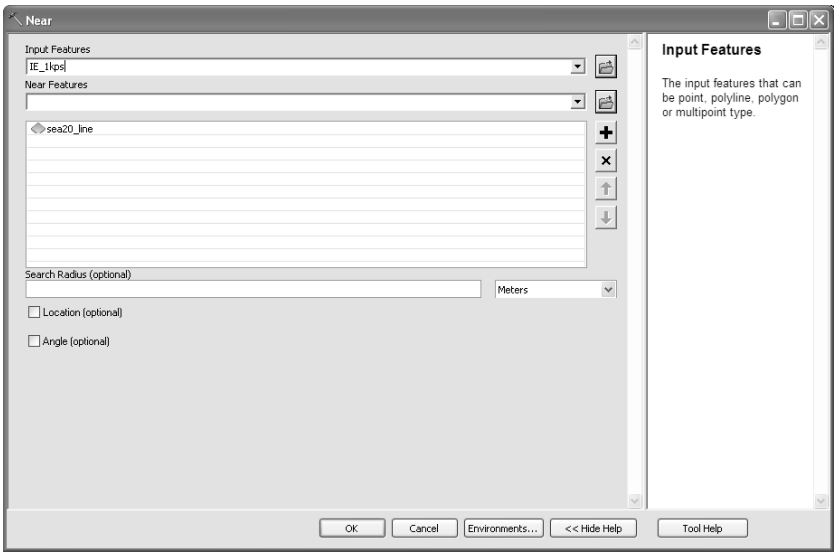
   4- Clean the buffer line because it was in both side by delete the inside line using customize toolbar\Editor\start editing\select the line to make a split using split tool.



B. generate and calculate the point distance to sea for each 1km grid:
    1. Using the 1km grid for each country for example I started with IE_1knew points were generated from the representative locations of input features (1km grid) saved it as IE_1kp.shp

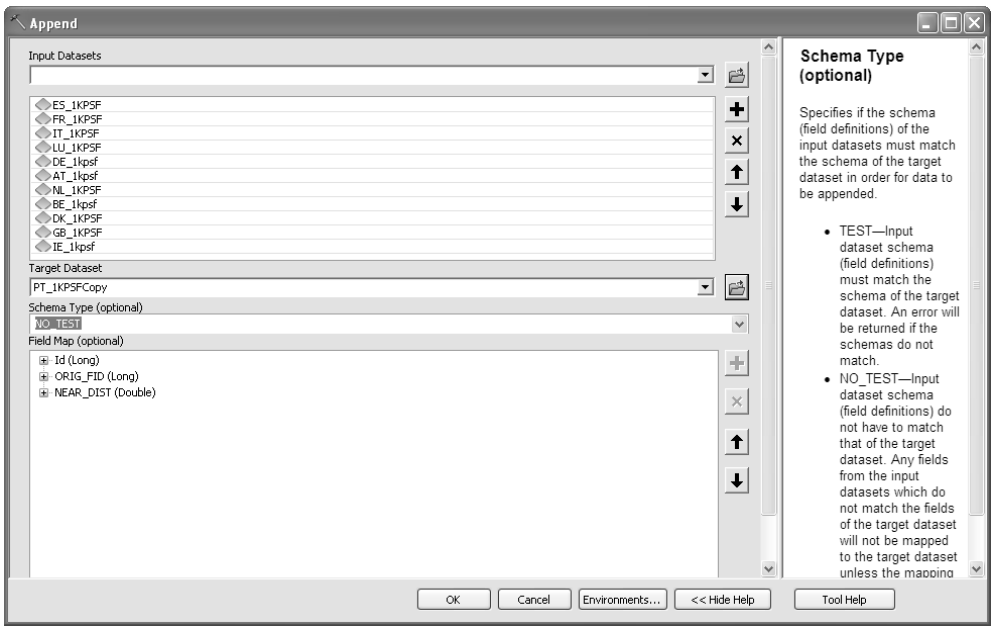


    2. To make the 1km grid math the 100mgrid cells the points coordination's(X,Y) have to be shifted by:
        i. Add new fields (type:Double) and calculate (x_corrdinate) and (y_corrdinate).
        ii. Add new field (shift_x) and (shift_y) which =x_cordinate -50 same with y.
        iii. Then export the table IE_1kps.dbf, and then remove it from workspace.
        iv. Add again the IE_1kps.dbf and display the x, Y coordinates.
        v. It will be add as an event therefore it has to be exported and saved as IE_1kpsf.shp
    3. Calculate the distance to sea for each point using NEAR command.



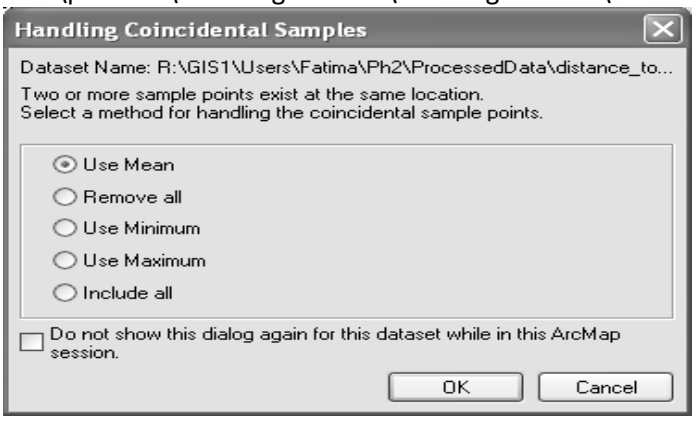
C. To clean the attribute table first keep the field for Fid, Shape and Near_Dist in the chosen country which will use it as a centre to append the rest of the countries.

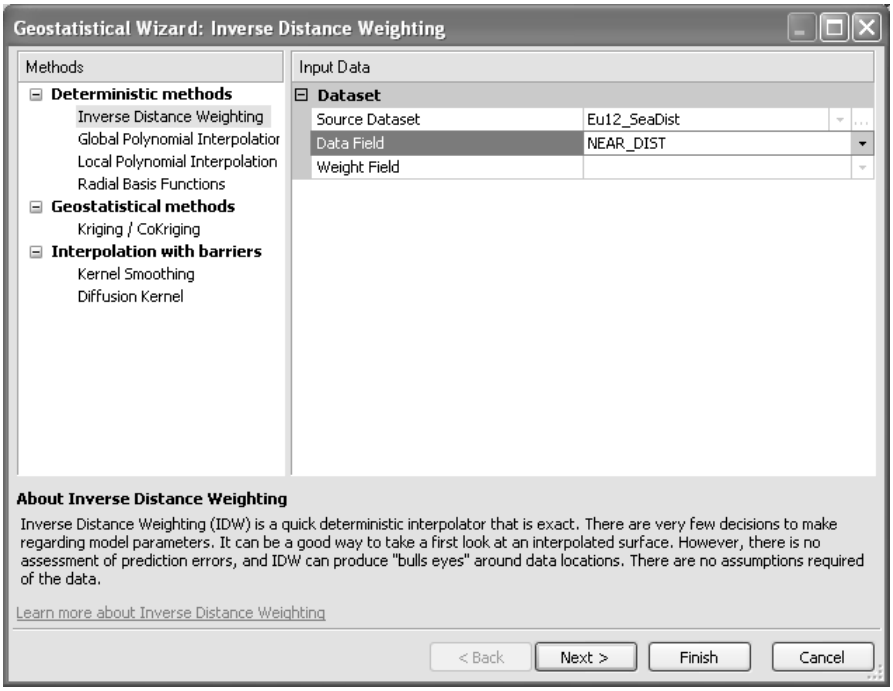
1. Use Delete field/field/data management tools command
2. Use Append/ General/Data management tools command to append all countries together.

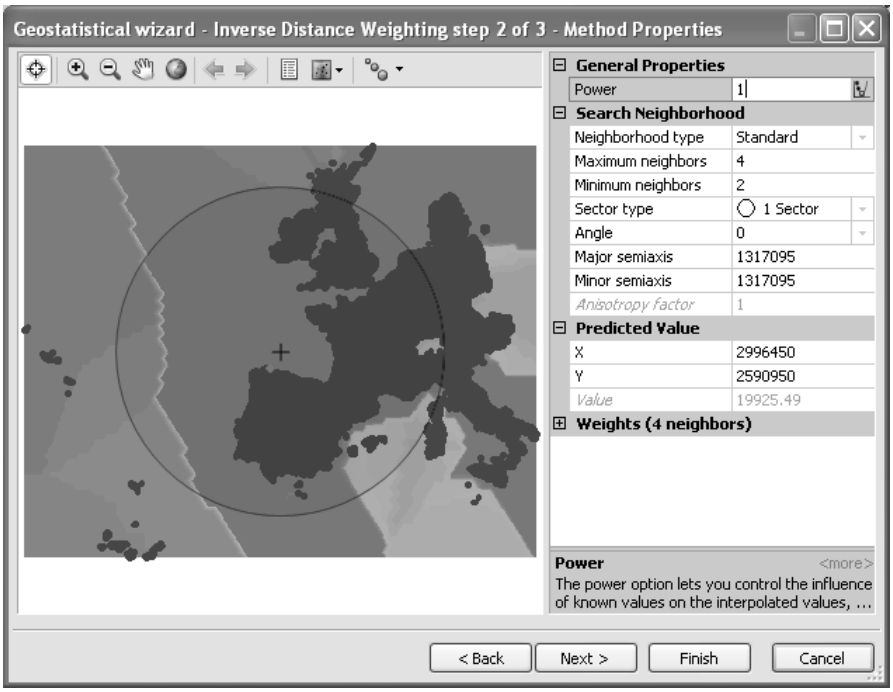


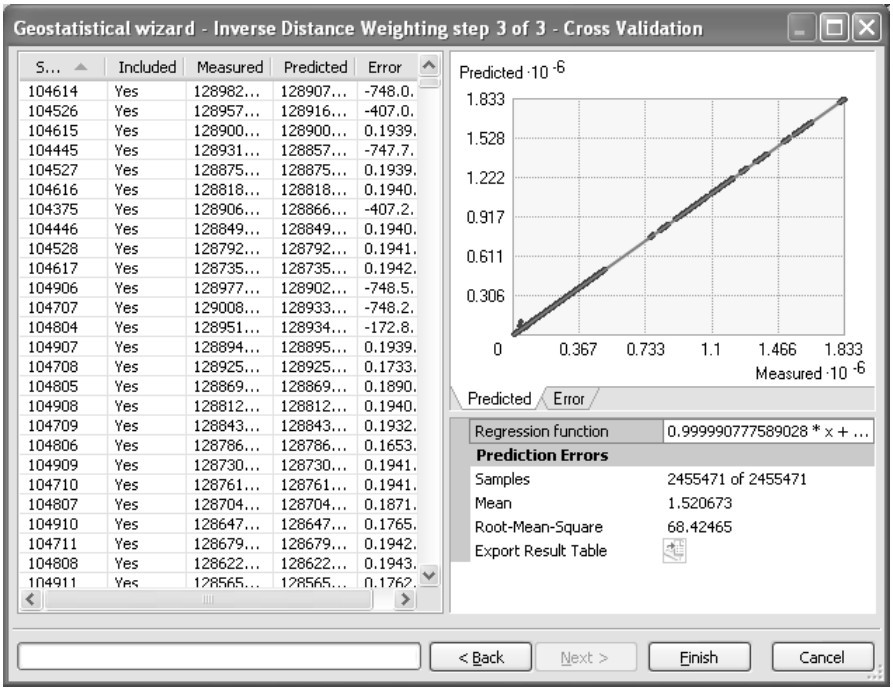
3. Rename the target dataset to E12_SEaDist.shp

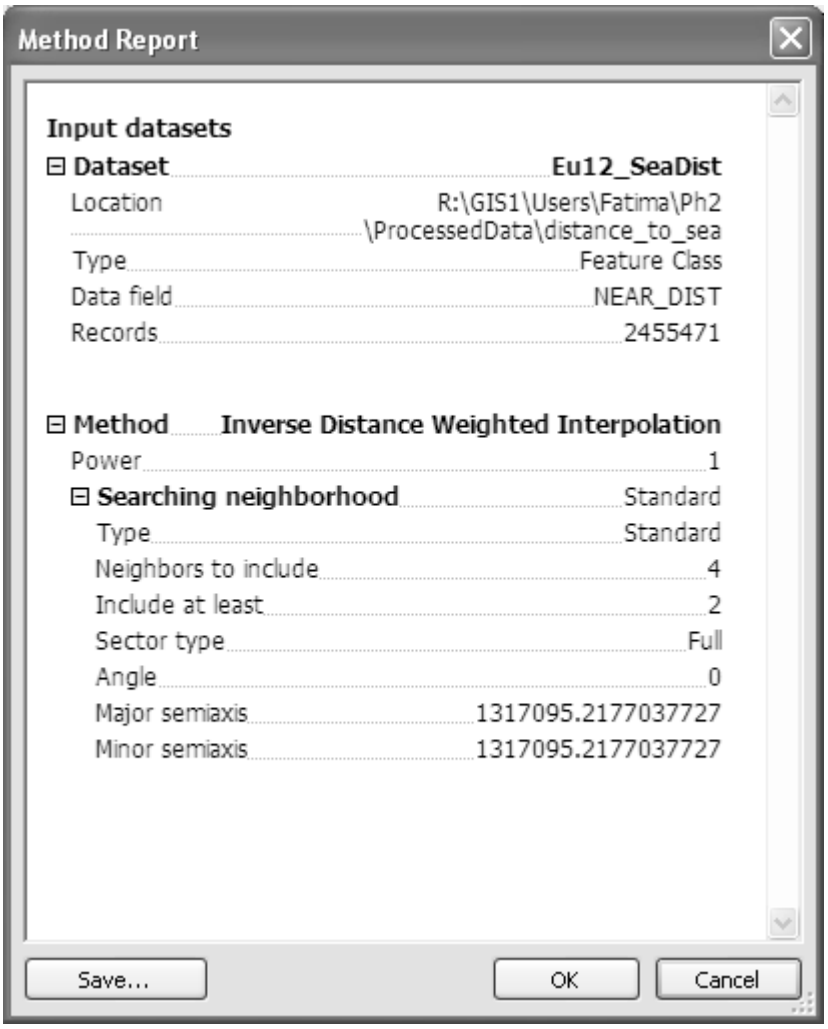
D. Interpolation:

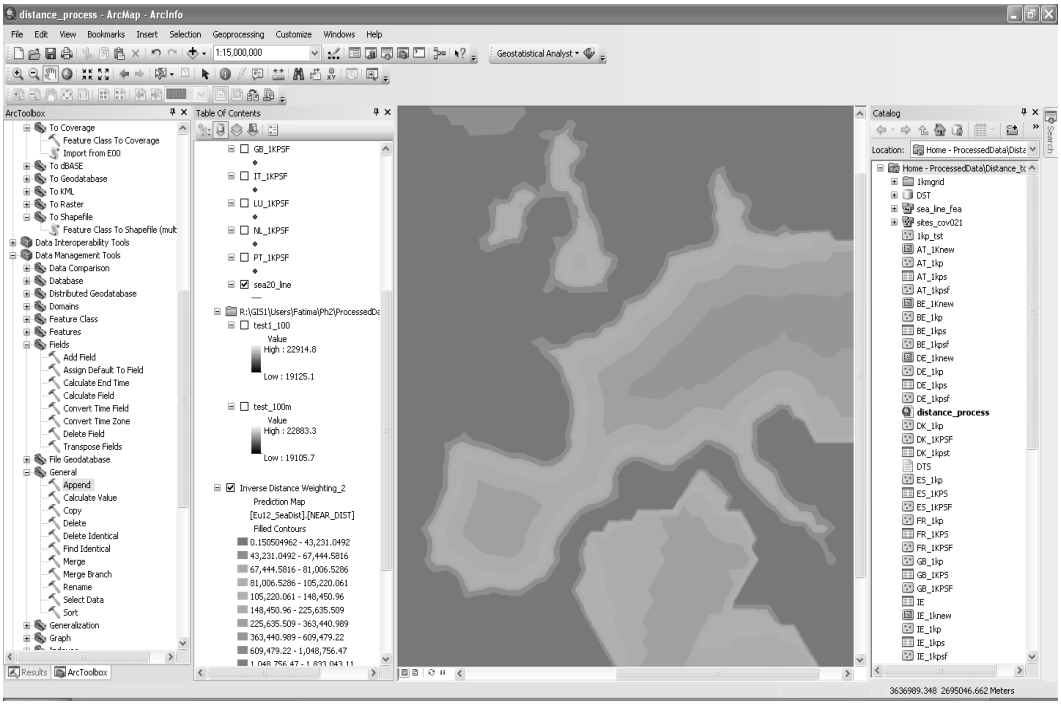
1. From customize select toolbar\geostatistic analyst\wizard\deterministic method\inverse distance weighting\sourcepoint: IE_1kpsf.shp\Data field: near field\power:1\max neighbour 4\min neighbour 2\ one sector.
2. 


Geostatistical Wizard: Inverse Distance Weighting
Methods
Deterministic methods
Inverse Distance Weighting
Global Polynomial Interpolatior
Local Polynomial Interpolation
Radial Basis Functions
Geostatistical methods
Kriging / CoKriging
Interpolation with barriers
Kernel Smoothing
Diffusion Kernel
Input Data
Dataset
Source Dataset
Eu12_SeaDist
Data Field
NEAR_DIST
Weight Field
About Inverse Distance Weighting
Inverse Distance Weighting (IDW) is a quick deterministic interpolator that is exact. There are very few decisions to make regarding model parameters. It can be a good way to take a first look at an interpolated surface. However, there is no assessment of prediction errors, and IDW can produce "bulls eyes" around data locations. There are no assumptions required of the data.
Learn more about Inverse Distance Weighting
< Back
Next >
Finish
Cancel

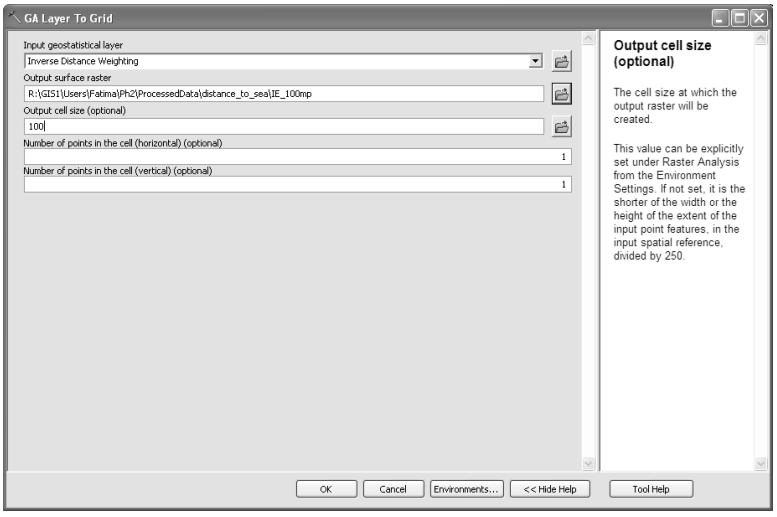

Geostatistical wizard - Inverse Distance Weighting step 2 of 3 - Method Properties
General Properties
Power
1
Search Neighborhood
Neighborhood type
Standard
Maximum neighbors
4
Minimum neighbors
2
Sector type
1 Sector
Angle
0
Major semiaxis
1317095
Minor semiaxis
1317095
Anisotropy factor
1
Predicted Value
X
2996450
Y
2590950
Value
19925.49
Weights (4 neighbors)
Power
<more>
The power option lets you control the influence of known values on the interpolated values, ...
< Back
Next >
Finish
Cancel


Geostatistical wizard - Inverse Distance Weighting step 3 of 3 - Cross Validation
S... Included Measured Predicted Error
104614 Yes 128982... 128907... -748.0.
104526 Yes 128957... 128916... -407.0.
104615 Yes 128900... 128900... 0.1939.
104445 Yes 128931... 128857... -747.7.
104527 Yes 128875... 128875... 0.1939.
104616 Yes 128818... 128818... 0.1940.
104375 Yes 128906... 128866... -407.2.
104446 Yes 128849... 128849... 0.1940.
104528 Yes 128792... 128792... 0.1941.
104617 Yes 128735... 128735... 0.1942.
104906 Yes 128977... 128902... -748.5.
104707 Yes 129008... 128933... -748.2.
104804 Yes 128951... 128934... -172.8.
104907 Yes 128894... 128895... 0.1939.
104708 Yes 128925... 128925... 0.1733.
104805 Yes 128869... 128869... 0.1890.
104908 Yes 128812... 128812... 0.1940.
104709 Yes 128843... 128843... 0.1932.
104806 Yes 128786... 128786... 0.1653.
104909 Yes 128730... 128730... 0.1941.
104710 Yes 128761... 128761... 0.1941.
104807 Yes 128704... 128704... 0.1871.
104910 Yes 128647... 128647... 0.1765.
104711 Yes 128679... 128679... 0.1942.
104808 Yes 128622... 128622... 0.1943.
104911 Yes 128565... 128565... 0.1762.
Predicted ·10 -6
1.833
1.528
1.222
0.917
0.611
0.306
0
0.367
0.733
1.1
1.466
1.833
Measured ·10 -6
Predicted
Error
Regression function
0.999990777589028 * x + ...
Prediction Errors
Samples
2455471 of 2455471
Mean
1.520673
Root-Mean-Square
68.42465
Export Result Table
< Back
Next >
Finish
Cancel

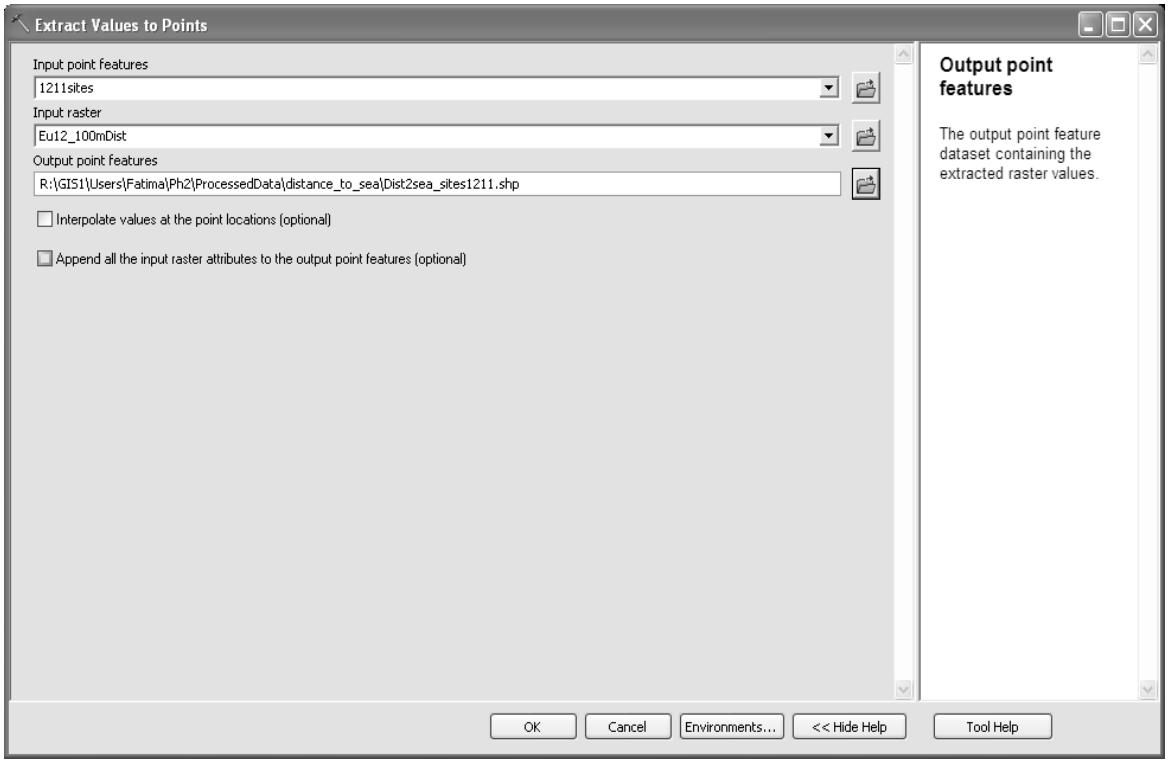

Method Report
Input datasets
Dataset
Eu12_SeaDist
Location
R:\GIS1\Users\Fatima\Ph2\ProcessedData\distance_to_sea
Type
Feature Class
Data field
NEAR_DIST
Records
2455471
Method
Inverse Distance Weighted Interpolation
Power
1
Searching neighborhood
Standard
Type
Standard
Neighbors to include
4
Include at least
2
Sector type
Full
Angle
0
Major semiaxis
1317095.2177037727
Minor semiaxis
1317095.2177037727
Save...
OK
Cancel



3. Right click on the resulted interpolation file to export data as a raster



4. Extract the sites distance from the sea by using Spatial Analyst Tools/Extraction/Extraction values to points



For validation the distance to sea for the sites point was calculated and compared with distance to sea from the interpolation intersected with resulting raster after interpolation.

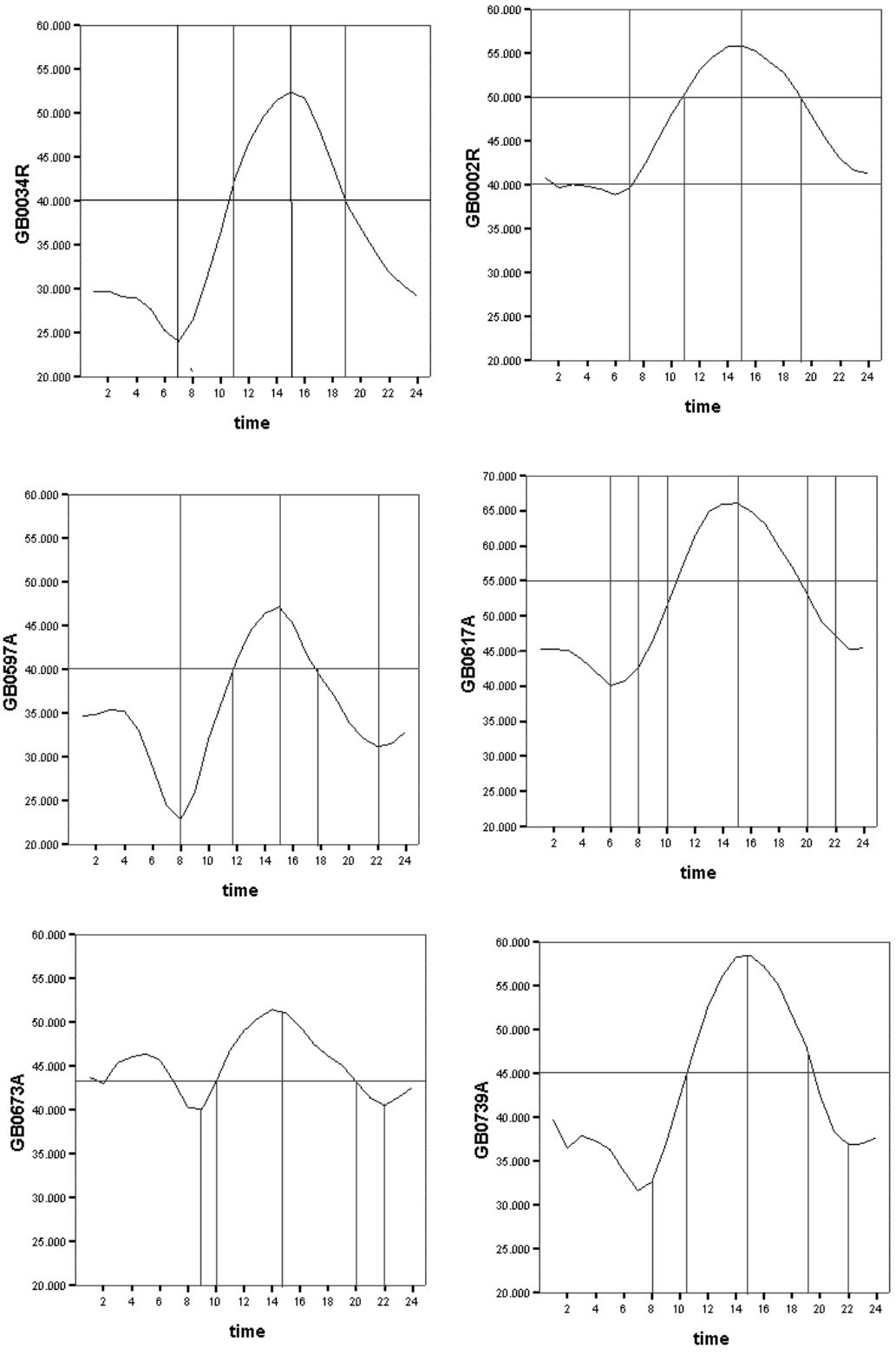
Correlations

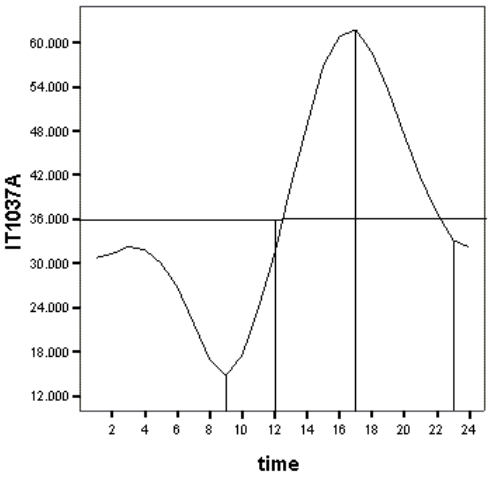
		RASTERVALU
DTS	Pearson Correlation	.99(**)
	Sig. (2-tailed)	.000
	N	1211

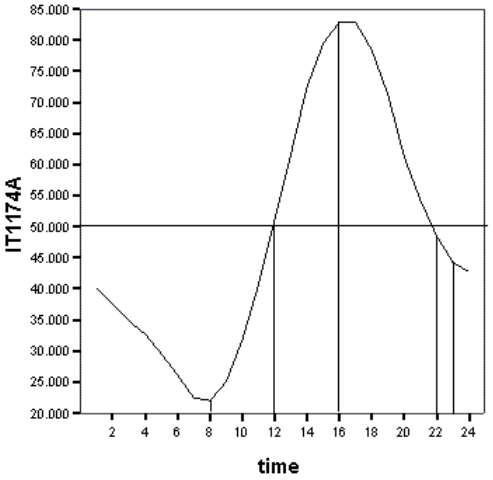
** Correlation is significant at the 0.01 level (2-tailed).

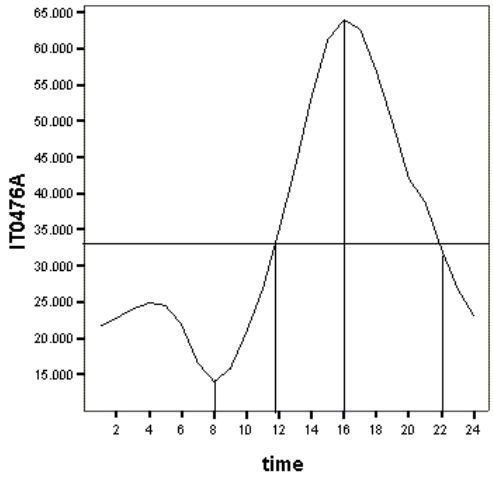
V Define the critical time period during the day

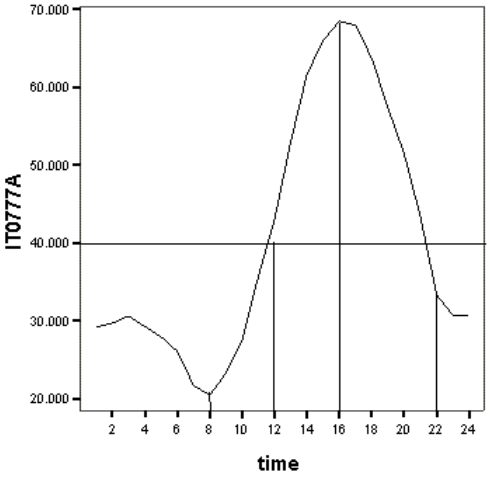
FigureA.1: line charts summarizes O_3 concentration over Time for a selection of sites

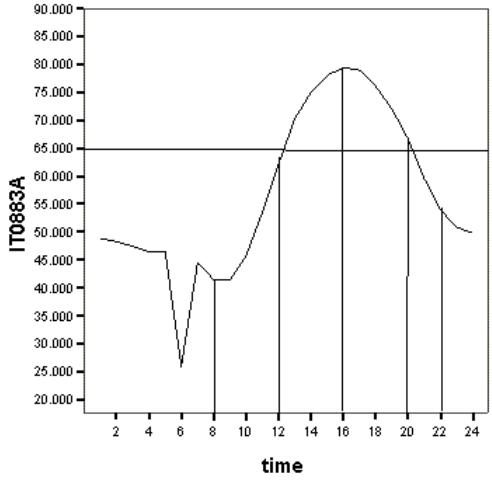


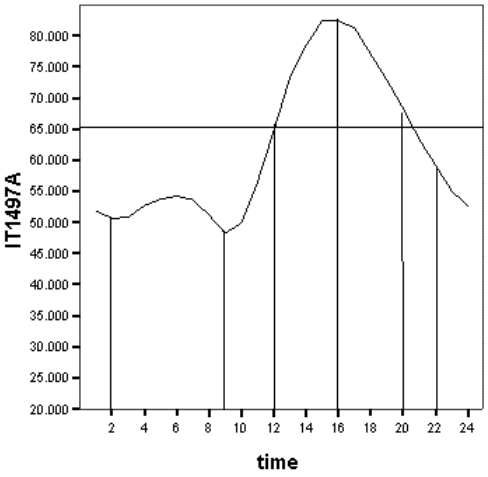

IT1037A
60.000
54.000
48.000
42.000
36.000
30.000
24.000
18.000
12.000
2 4 6 8 10 12 14 16 18 20 22 24
time


IT1174A
85.000
80.000
75.000
70.000
65.000
60.000
55.000
50.000
45.000
40.000
35.000
30.000
25.000
20.000
2 4 6 8 10 12 14 16 18 20 22 24
time


IT0476A
65.000
60.000
55.000
50.000
45.000
40.000
35.000
30.000
25.000
20.000
15.000
2 4 6 8 10 12 14 16 18 20 22 24
time


IT0777A
70.000
60.000
50.000
40.000
30.000
20.000
2 4 6 8 10 12 14 16 18 20 22 24
time


IT0883A
90.000
85.000
80.000
75.000
70.000
65.000
60.000
55.000
50.000
45.000
40.000
35.000
30.000
25.000
20.000
2 4 6 8 10 12 14 16 18 20 22 24
time


IT1497A
80.000
75.000
70.000
65.000
60.000
55.000
50.000
45.000
40.000
35.000
30.000
25.000
20.000
2 4 6 8 10 12 14 16 18 20 22 24
time

VI. The ratio of sites per 10,000 km^2 for different site types in different countries

Table A.2 Density of sites in each country of study area by the thirteen Site types

EU country	Area (km^2)	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
AT	83,859	.24	3.22	0.72	2.98	0.12	.24	.12	.00	.00	2.03	.00	2.62	0.48	12.76
BE	30,518	.98	0.00	1.97	0.00	0.00	.00	.00	.98	7.21	0.00	.00	.00	0.00	11.14
DK	43,094	.00	0.70	0.00	0.00	0.00	.00	.46	.23	0.00	0.00	.00	.00	0.00	1.39
ES	504,790	.06	0.71	0.44	0.14	0.14	.40	.57	.67	0.02	0.04	.91	.48	0.02	4.60
FR	543,965	.44	0.74	1.12	0.97	1.08	.26	.11	.33	0.68	0.35	.07	.29	0.06	6.51
GB	243,820	.08	0.00	0.08	0.00	0.04	.00	.74	.90	0.33	0.00	.00	.41	0.00	2.58
GE	357,022	.84	0.36	1.20	0.76	1.88	.34	.22	.42	0.95	0.53	.00	.39	0.00	7.90
IE	70,723	.00	0.00	0.00	0.00	0.00	.00	.00	.14	0.00	0.00	.00	.71	0.00	0.85
IT	301,316	.00	0.33	0.00	0.10	0.00	.27	.00	.03	0.00	0.20	.17	.13	1.16	2.39
NL	41,526	.00	0.24	0.24	0.24	1.93	.00	.24	.48	3.85	0.00	.00	.00	0.00	7.22
PT	91,906	.33	0.76	0.22	0.00	0.65	.22	.00	.44	0.11	0.00	.00	.00	0.00	2.72

VII. Perl scripts

Compute_Ozone

```
#! /usr/local/bin/perl –w #Feb 2012 #create ozone estimates for Fatima, step 2 #LUR_grids.csv: Point_ID, LUR #Met_grid.csv: input_FID (1-4), value per day #grid_total.csv:TF estimates by point_id and day #tf_fid.csv: Point, FID, TF*0.747 for each day # TF * 0.747 + LUR based on Point_ID + Met based on FID and day -23.9 use strict; my $tf_fid = "/home/EPH/archive/gds_data/gis/temp/tf_fid.csv"; my $lur = "Lur_grids.csv"; my $met = "Met_grid.csv"; my $out_file = "/home/EPH/archive/gds_data/gis/temp/ozone_estimate.csv"; #first get Met/FID data into memory open MET, $met or die "Cannot open $met"; my %met_data; my $met_cols = <MET>; my @days = split ",", $met_cols; my $num_days = scalar(@days) - 1; if ($num_days != 2191) { die "Num days $num_days != 2191"; }while (my $line = <MET>) { if ($line !~ /Input_FID/) {chomp $line; my @vals = split ",", $line; for (my $i = 1; $i <= $num_days; $i++) { $met_data{$vals[0]}[$i-1]=$vals[$i];} }} close MET; open LUR, $lur or die "Cannot open $lur"; open TF, $tf_fid or die "Cannot open $tf_fid"; open OUT, ">$out_file" or die "Cannot open $out_file"; my $point = 0; my $linel = <LUR>; while (my $linet = <TF>) { $linel = <LUR>; chomp $linet; chomp $linel; my @tf_vals = split ",", $linet; my @lur_vals = split ",", $linel; if ($tf_vals[0] != $point) { die "$point != TF $tf_vals[0]"; } elsif ($lur_vals[0] != $point) { die "$point != LUR $lur_vals[0]"; } elsif (scalar(@tf_vals) != ($num_days + 2)) { die "TF ".scalar(@tf_vals)." != $num_days+2"; print OUT "$point"; my $fid = $tf_vals[1]; if ($fid <0 or $fid > 4) { die "Invalid $fid FID for $point"; } for (my $i=0; $i < $num_days; $i++) { my $sum = -23.9 + $lur_vals[1] + $met_data{$fid}[$i] + $tf_vals[$i+2];printf OUT ",%.2f", $sum; } print OUT "\n"; $point++; } close OUT; close TF; close LUR;
```

Extract _point _data

```
#! /usr/local/bin/perl –w #Feb 2012 #Fatima has a list of Points she would like to see the data for use strict;my$point_file="fatima_Nl_point.txt";my$ozone_file="/home/EPH/archive/margaret_data/ozone_estimate.csv"; my $out_file = "/home/EPH/archive/margaret_data/Nl_points_ozone.csv"; my @point_ids; open PT, $point_file or die "Cannot open $point_file"; while (my $line = <PT>) { if ($line =~ /(\d+)/) { push @point_ids, $1; }} close PT; my $day_file = "days.csv";open DAY, $day_file or die "Cannot open $day_file"; my $day_line = <DAY>; close DAY; my $p = 0; open OUT, ">$out_file" or die "Cannot open $out_file"; open OZ, $ozone_file or die "Cannot open $ozone_file"; print OUT "Point_id,$day_line"; while (my $line = <OZ>) { if ($line =~ /^(\d+),/) { my $pid = $1; if ($pid == $point_ids[$p]) { print OUT $line; $p++;} }} close OZ; close OUT;
```

Appendix B

Analyses output

Contents

I ANOVA outcome results of the thirteen site type in study area countries

Table B.1 ANOVA for Classification Sites

country		Sum of Squares	df	Mean Square	F	Sig.	
AT	Between Groups	25665.62	9	2851.735	64.346	0	86%
	Within Groups	4298.919	97	44.319			
	Total	29964.54	106				
BE	Between Groups	1814.882	3	604.961	23.084	0	70%
	Within Groups	786.206	30	26.207			
	Total	2601.087	33				
DE	Between Groups	23772.44	10	2377.244	85.489	0	76%
	Within Groups	7535.877	271	27.808			
	Total	31308.31	281				
DK	Between Groups	653.099	2	326.549	43.615	0.006	97%
	Within Groups	22.461	3	7.487			
	Total	675.56	5				
ES	Between Groups	39836.06	12	3319.671	39.74	0	69%
	Within Groups	18294.19	219	83.535			
	Total	58130.25	231				
FR	Between Groups	24098.78	12	2008.231	52.699	0	65%
	Within Groups	12994.79	341	38.108			
	Total	37093.56	353				
GB	Between Groups	5419.738	6	903.29	24.203	0	72%
	Within Groups	2089.965	56	37.321			
	Total	7509.703	62				
IE	Between Groups	310.923	1	310.923	3.949	0.118	50%
	Within Groups	314.97	4	78.742			
	Total	625.893	5				
IT	Between Groups	9199.245	7	1314.178	17.992	0	66%
	Within Groups	4674.765	64	73.043			
	Total	13874.01	71				
NL	Between Groups	589.161	6	98.194	6.542	0	63%
	Within Groups	345.227	23	15.01			
	Total	934.389	29				
PT	Between Groups	1551.963	6	258.661	8.622	0	74%
	Within Groups	539.998	18	30			
	Total	2091.961	24				

II PCA output

Table B.2 depicts the correlation matrix for the 21 indicators. Inspection of the correlation matrix reveals that 184 of the 209 correlations (89%) are significant at the .01 level. This provides an adequate to proceeding to the next level which is assessing the overall significance of the correlation matrix with the Bartlett test. In this analysis the overall correlation are significant at the 0.0001 level, also the Measure of sampling adequacy(KMO) equal 0.889, furthermore, each indicators exceed the thresholds value (0.5), Table B.3 These measures all indicate that the reduced set of indicators is appropriate for factor analysis, and analysis can proceed to the next stages.

Table B.2 the correlation matrix between the 21 indicators

	SUM_Nmean	WINT_Nmean	WD_Nmean	WE_Nmean	AM_Nmean	PM_Nmean	NIGHT_Nmean	SUM_Nvar	WINT_Nvar	WD_Nvar	WE_Nvar	AM_Nvar	PM_Nvar	NIGHT_Nvar	SUM_Nmax	WINT_Nmax	WD_Nmax	WE_Nmax	AM_Nmax	PM_Nmax
SUM_Nmean	1.00																			
WINT_Nmean	-0.93	1.00																		
WD_Nmean	-0.42	0.40	1.00																	
WE_Nmean	0.42	-0.40	-1.00	1.00																
AM_Nmean	-0.43	0.47	0.27	-0.27	1.00															
PM_Nmean	0.65	-0.69	-0.46	0.46	-0.44	1.00														
NIGHT_Nmean	-0.47	0.47	0.26	-0.26	-0.01	-0.86	1.00													
SUM_Nvar	0.19	-0.14	-0.46	0.46	-0.11	0.21	-0.18	1.00												
WINT_Nvar	0.62	-0.64	-0.65	0.65	-0.36	0.68	-0.53	0.52	1.00											
WD_Nvar	0.82	-0.80	-0.68	0.68	-0.46	0.72	-0.52	0.57	0.88	1.00										
WE_Nvar	0.81	-0.81	-0.63	0.63	-0.51	0.71	-0.47	0.53	0.86	0.98	1.00									
AM_Nvar	0.73	-0.77	-0.70	0.70	-0.54	0.85	-0.61	0.42	0.85	0.92	0.92	1.00								
PM_Nvar	0.73	-0.68	-0.78	0.78	-0.34	0.61	-0.42	0.61	0.86	0.95	0.93	0.86	1.00							
NIGHT_Nvar	0.62	-0.66	-0.65	0.65	-0.32	0.87	-0.76	0.43	0.81	0.85	0.83	0.92	0.79	1.00						
SUM_Nmax	0.07	0.00	-0.44	0.44	0.03	0.09	-0.13	0.85	0.43	0.43	0.36	0.28	0.51	0.32	1.00					
WINT_Nmax	0.66	-0.73	-0.59	0.59	-0.36	0.72	-0.57	0.44	0.91	0.87	0.85	0.87	0.81	0.84	0.35	1.00				
WD_Nmax	0.33	-0.34	-0.78	0.77	-0.23	0.52	-0.40	0.62	0.79	0.76	0.72	0.75	0.82	0.78	0.62	0.72	1.00			
WE_Nmax	0.27	-0.29	-0.71	0.71	-0.26	0.46	-0.32	0.57	0.76	0.71	0.70	0.71	0.77	0.72	0.56	0.68	0.95	1.00		
AM_Nmax	0.63	-0.71	-0.62	0.62	-0.56	0.90	-0.65	0.25	0.72	0.76	0.77	0.94	0.69	0.86	0.12	0.77	0.62	0.59	1.00	
PM_Nmax	0.33	-0.42	-0.73	0.73	-0.10	0.41	-0.24	0.33	0.58	0.58	0.60	0.67	0.69	0.65	0.26	0.58	0.69	0.68	0.64	1.00

Table B.3 Assessing the appropriate of PCA by Measure of Sampling Adequacy (KMO)

Indicators	SUM_Nmean	WINT_Nmean	WD_Nmean	WE_Nmean	AM_Nmean	PM_Nmean	NIGHT_Nmean	SUM_Nvar	WINT_Nvar	WD_Nvar	WE_Nvar	AM_Nvar	PM_Nvar	NIGHT_Nvar	SUM_Nmax	WINT_Nmax	WD_Nmax	WE_Nmax	AM_Nmax	PM_Nmax	NIGHT_Nmax
SUM_Nmean	0.90																				
WINT_Nmean	0.38	0.92																			
WD_Nmean	0.07	0.06	0.88																		
WE_Nmean	0.06	0.06	1.00	0.88																	
AM_Nmean	-0.13	0.04	0.04	0.04	0.60																
PM_Nmean	-0.04	0.01	-0.03	-0.04	0.56	0.87															
NIGHT_Nmean	-0.06	0.03	0.03	0.02	0.78	0.79	0.75														
SUM_Nvar	0.40	-0.20	0.03	0.03	-0.06	0.04	0.00	0.82													
WINT_Nvar	0.15	-0.15	-0.01	-0.01	0.04	0.00	0.11	0.11	0.95												
WD_Nvar	-0.29	0.32	0.00	-0.01	0.24	0.17	0.26	-0.26	-0.06	0.91											
WE_Nvar	-0.07	0.13	-0.01	0.01	0.20	-0.12	-0.11	-0.18	-0.14	-0.23	0.93										
AM_Nvar	0.13	-0.03	-0.08	-0.09	-0.17	0.11	-0.10	0.01	0.00	-0.37	-0.30	0.91									
PM_Nvar	-0.27	-0.28	0.03	0.02	-0.21	-0.15	-0.09	-0.04	-0.12	-0.46	-0.25	0.09	0.92								
NIGHT_Nvar	0.02	-0.17	0.04	0.03	-0.18	-0.25	-0.11	0.03	0.10	-0.30	-0.34	0.00	0.44	0.90							
SUM_Nmax	-0.17	-0.03	-0.08	-0.08	0.02	0.06	0.05	-0.69	0.03	-0.03	0.10	0.18	-0.01	0.03	0.82						
WINT_Nmax	0.07	0.30	0.08	0.08	-0.13	0.03	-0.09	0.02	-0.64	-0.07	0.05	-0.02	0.10	-0.03	-0.10	0.94					
WD_Nmax	0.24	-0.07	-0.01	-0.02	-0.03	-0.04	-0.06	0.21	0.01	-0.27	0.21	-0.03	-0.07	-0.13	-0.28	-0.03	0.93				
WE_Nmax	0.16	-0.18	0.10	0.11	0.07	0.06	-0.01	0.19	-0.03	0.06	-0.20	-0.07	-0.11	0.08	-0.05	-0.02	-0.60	0.93			
AM_Nmax	-0.14	0.05	0.10	0.10	0.32	-0.32	0.06	-0.06	0.01	0.26	0.27	-0.82	0.03	0.11	-0.12	-0.07	0.03	0.04	0.87		
PM_Nmax	0.25	0.23	-0.06	-0.07	-0.58	-0.03	-0.40	0.04	0.04	0.16	-0.29	0.28	-0.30	-0.11	0.05	0.06	-0.05	-0.01	-0.52	0.84	
NIGHT_Nmax	-0.04	0.12	-0.01	-0.02	0.25	0.26	0.40	-0.05	0.03	0.31	0.25	-0.15	-0.26	-0.76	0.02	-0.09	0.00	-0.14	-0.02	-0.13	0.89

Overall KMO: 0.9

III MLOR output

1. Selecting the predictor variables

Case Processing Summary

		N	Marginal Percentage
Site type	1	67	5.5%
	2	137	11.3%
	3	143	11.8%
	4	116	9.6%
	5	149	12.3%
	6	58	4.8%
	7	65	5.4%
	8	101	8.3%
	9	119	9.8%
	10	63	5.2%
	11	55	4.5%
	12	95	7.8%
	13	43	3.6%
Valid		1211	100.0%
Missing		0	
Total		1211	
Subpopulation		1211a	

2. stepwise analysis, in series of stages

a. Major road variables

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_100 mr_500 mr_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=mr_10000 mr_100 mr_500
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5863.396	409.455	12	.000
mr_10000	5816.099	362.158	12	.000
mr_500	5481.016	27.076	12	.008

b. Secondary road variables:

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_500 mr_10000 sr_100 sr_500 sr_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)/MODEL=| FSTEP=mr_10000 mr_500 sr_100 sr_500 sr_10000
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE/PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB.
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5875.450	482.264	12	.000
mr_10000	5536.383	143.198	12	.000
mr_500	5425.635	32.449	12	.001
sr_100	5415.467	22.282	12	.034
sr_500	5424.553	31.367	12	.002
sr_10000	5473.754	80.569	12	.000

c. Local road variables:

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_500 mr_10000 sr_100 sr_500 sr_10000 lr_100
lr_500 lr_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=mr_10000 mr_500 sr_100 sr_500 sr_10000 lr_100 lr_500 lr_10000
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB.
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5759.075	637.091	12	.000
mr 10000	5216.530	94.546	12	.000
mr_500	5163.992	42.008	12	.000
sr_100	5143.645	21.661	12	.042
sr_500	5166.044	44.060	12	.000
sr 10000	5169.561	47.577	12	.000
lr 500	5234.892	112.908	12	.000
lr_10000	5224.466	102.482	12	.000

d. High density residential variables

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_500 mr_10000 sr_100 sr_500 sr_10000 lr_500
lr_10000 highdr_500 highdr_1000 highdr_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=mr_10000 mr_500 sr_100 sr_500 sr_10000 lr_500 lr_10000 highdr_500 highdr_1000
highdr_10000
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB.
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5515.937	561.502	12	.000
mr 10000	5053.517	99.082	12	.000
mr 500	4995.413	40.978	12	.000
sr_100	4975.538	21.103	12	.049
sr_500	4982.218	27.783	12	.006
sr_10000	5017.843	63.407	12	.000
lr 500	5006.130	51.695	12	.000
lr 10000	5055.183	100.748	12	.000
highdr_1000	5005.796	51.360	12	.000
highdr_10000	5046.868	92.433	12	.000

e. Low density residential variables

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_500 mr_10000 sr_100 sr_500 sr_10000 lr_500
lr_10000 highdr_1000 highdr_10000 lowdr_1000 lowdr_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)/MODEL=|  FSTEP=mr_10000 mr_500 sr_100 sr_500 sr_10000 lr_500 lr_10000
highdr_1000 highdr_10000 lowdr_1000 lowdr_10000
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE/PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI/SAVE PREDCAT PCPROB
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5288.774	465.957	12	.000
mr 10000	4877.492	54.675	12	.000
sr 100	4850.375	27.558	12	.006
sr_10000	4882.510	59.693	12	.000
lr_500	4848.227	25.410	12	.013
lr_10000	4910.042	87.225	12	.000
highdr_1000	4898.047	75.230	12	.000
highdr 10000	4875.825	53.008	12	.000
lowdr_1000	4951.039	128.222	12	.000
lowdr_10000	4872.513	49.695	12	.000

f. Industrial and commercial variables

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_10000 sr_100 sr_10000 lr_500 lr_10000
highdr_1000 highdr_10000 lowdr_1000 lowdr_10000 Ind/com_1000 Ind/com_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=mr_10000 sr_100 sr_10000 lr_500 lr_10000 highdr_1000 highdr_10000 lowdr_1000
lowdr_10000 Ind/com_1000 Ind/com_10000
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB.
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5167.492	416.816	12	.000
mr 10000	4800.034	49.358	12	.000
sr 100	4779.164	28.488	12	.005
sr_10000	4808.722	58.046	12	.000
lr_500	4773.362	22.686	12	.031
lr_10000	4831.105	80.429	12	.000
highdr 1000	4820.133	69.457	12	.000
highdr 10000	4817.010	66.334	12	.000
lowdr 1000	4871.966	121.290	12	.000
lowdr_10000	4795.771	45.095	12	.000
Ind/com_1000	4776.139	25.463	12	.013
Ind/com_10000	4784.470	33.794	12	.001

g. Forest variables

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_10000 sr_100 sr_10000 lr_500 lr_10000
highdr_1000 highdr_10000 lowdr_1000 lowdr_10000 nres_1000 nres_10000 forest_1000 forest_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)/MODEL=| FSTEP=mr_10000 sr_100 sr_10000 lr_500 lr_10000 highdr_1000
highdr_10000 lowdr_1000 lowdr_10000 nres_1000 nres_10000 forest_1000 forest_10000
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	4847.309	276.758	12	.000
mr_10000	4616.386	45.835	12	.000
sr_100	4598.696	28.145	12	.005
sr 10000	4626.559	56.008	12	.000
lr 500	4591.888	21.337	12	.046
lr_10000	4651.308	80.757	12	.000
highdr_1000	4629.845	59.294	12	.000
highdr_10000	4633.670	63.118	12	.000
lowdr_1000	4650.757	80.206	12	.000
lowdr 10000	4617.191	46.640	12	.000
nres 1000	4593.219	22.668	12	.031
nres_10000	4598.222	27.671	12	.006
forest_1000	4609.352	38.801	12	.000
forest_10000	4679.099	108.548	12	.000

h. Green area (agriculture and herbs) variables

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_10000 sr_100 sr_10000 lr_500 lr_10000
highdr_1000 highdr_10000 lowdr_1000 lowdr_10000 nres_1000 nres_10000 forest_1000 forest_10000
agri_1000 agri_5000 herb_1000 herb_5000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=mr_10000 sr_100 sr_10000 lr_500 lr_10000 highdr_1000 highdr_10000 lowdr_1000
lowdr_10000 nres_1000 nres_10000 forest_1000 forest_10000 agri_1000 herb_1000 herb_5000
agri_5000/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)/INTERCEPT=INCLUDE /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB.
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	4548.337	103.545	12	.000
mr_10000	4481.626	36.834	12	.000
sr_100	4471.518	26.726	12	.008
sr_10000	4489.861	45.070	12	.000
lr_10000	4540.977	96.186	12	.000
highdr_1000	4529.732	84.940	12	.000
highdr_10000	4505.549	60.757	12	.000
lowdr_1000	4523.415	78.624	12	.000
lowdr_10000	4471.830	27.038	12	.008
Ids/com_1000	4472.159	27.367	12	.007
Ind/com_10000	4468.176	23.384	12	.025
forest_1000	4483.548	38.756	12	.000
forest_10000	4541.937	97.146	12	.000
herb_5000	4496.718	51.926	12	.000
agri_5000	4486.109	41.318	12	.000

I. Topography variables:

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH sr_100 sr_10000 lr_10000 ind/com_10000 highdr_1000
highdr_10000 lowdr_1000 lowdr_10000 forest_1000 forest_10000 herb_5000 agri_5000 Dis2sea
Altitude Topex ind/com_1000 mr_10000
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=sr_100 sr_10000 lr_10000 ind/com_10000 highdr_1000 highdr_10000 lowdr_1000
lowdr_10000 forest_1000 forest_10000 agri_5000 herb_5000 Altitude Dis2sea Topex ind/com_1000
mr_10000  /STEPWISE=PIN(.05)  POUT(0.051)  MINEFFECT(0)  RULE(SINGLE)  ENTRYMETHOD(LR)
REMOVALMETHOD(LR)/INTERCEPT=INCLUDE /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI/SAVE
PREDCAT.
```

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	4186.666	110.627	12	.000
sr_100	4100.311	24.272	12	.019
sr_10000	4120.867	44.828	12	.000
lr_10000	4181.911	105.871	12	.000
Ind/com_10000	4103.414	27.374	12	.007
highdr_1000	4151.423	75.384	12	.000
highdr_10000	4159.364	83.324	12	.000
lowdr_1000	4125.629	49.589	12	.000
forest_1000	4101.281	25.241	12	.014
forest_10000	4163.257	87.217	12	.000
agri_5000	4118.384	42.345	12	.000
herb_5000	4146.044	70.004	12	.000
Altitude	4288.146	212.107	12	.000
Dis2sea	4159.250	83.210	12	.000
Topex	4187.977	111.938	12	.000
Ind/com_1000	4098.724	22.684	12	.031
mr_10000	4111.491	35.451	12	.000

j. Meteorological factors variables:

```
NOMREG CLU13 (BASE=5 ORDER=ASCENDING) WITH mr_10000 sr_100 sr_10000 lr_10000 highdr_1000
highdr_10000 lowdr_1000 lowdr_10000 ind/com_1000 nres_10000 forest_1000 forest_10000 agri_5000
herb_5000 Altitude Topex Dis2sea tp_win tem_sum ws_win
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
  /MODEL=| FSTEP=mr_10000 sr_100 sr_10000 lr_10000 highdr_1000 highdr_10000 lowdr_1000
lowdr_10000 ind/com_1000 nres_10000 forest_1000 forest_10000 herb_5000 agri_5000 Topex ws_win
tp_win Altitude Dis2sea tem_sum
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
```

```
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE PREDCAT PCPROB.
```

Final Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	6034.4			
Final	3436.9	2597.5	228	.000

Pseudo R-Square

Cox and Snell	.9
Nagelkerke	.9
McFadden	.5

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	3577.235	140.302	12	.000
mr_10000	3463.796	26.862	12	.008
sr_100	3465.744	28.810	12	.004
sr 10000	3481.180	44.246	12	.000
lr 10000	3488.090	51.157	12	.000
highdr 1000	3497.529	60.595	12	.000
highdr_10000	3468.235	31.302	12	.002
lowdr_1000	3483.183	46.249	12	.000
lowdr_10000	3460.115	23.181	12	.026
Ind/com 1000	3462.087	25.153	12	.014
forest 1000	3462.591	25.657	12	.012
forest_10000	3485.150	48.216	12	.000
herb_5000	3476.889	39.955	12	.000
agri_5000	3474.227	37.293	12	.000
Topex	3510.303	73.370	12	.000
ws win	3798.004	361.071	12	.000
tp win	3543.831	106.898	12	.000
Altitude	3662.180	225.246	12	.000
Dis2sea	3513.051	76.118	12	.000
tem_sum	3564.733	127.799	12	.000

3. Explore the VIF

Variables	VIF
mr_10000	4.6
sr_10000	3.6
lr_10000	3.7
Ind/com_10000	4.2
lowdr_1000	1.9
highdr_1000	1.8
highdr_10000	4.2
sr_100	1.0
lowdr_10000	4.4
Ind/com_1000	1.3
herb_5000	2.2
agri_5000	3.5
forest_1000	2.3
forest_10000	3.5
Dis2sea	1.6
Altitude	2.2
Topex	1.2
ws_win	2.1
tp_win	1.6
tem_sum	2.1

Exclude variable if VIF >5

4. Kappa Indix

Symmetric Measures

		Value	Asymp. Std. Errora	Approx. Tb	Approx. Sig.
Measure of Agreement	Kappa	.5	.01	51.4	.000
N of Valid Cases		1211			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis

IV Training and validation datasets

From the results of ANOVA demonstrate that the means concentration between the two dataset was not significantly different, the t-statistic was non-significant p-value=0.5. Leven's test of equality of variance was not significant P-value=0.3, indicating homogeneity of variances of the two data sets. The same results in the box plot in Figure B.1 show that the two datasets are similar and also the descriptive statistics of 25, 50 and 75 percentile show the similarity as shown in Table B.4.

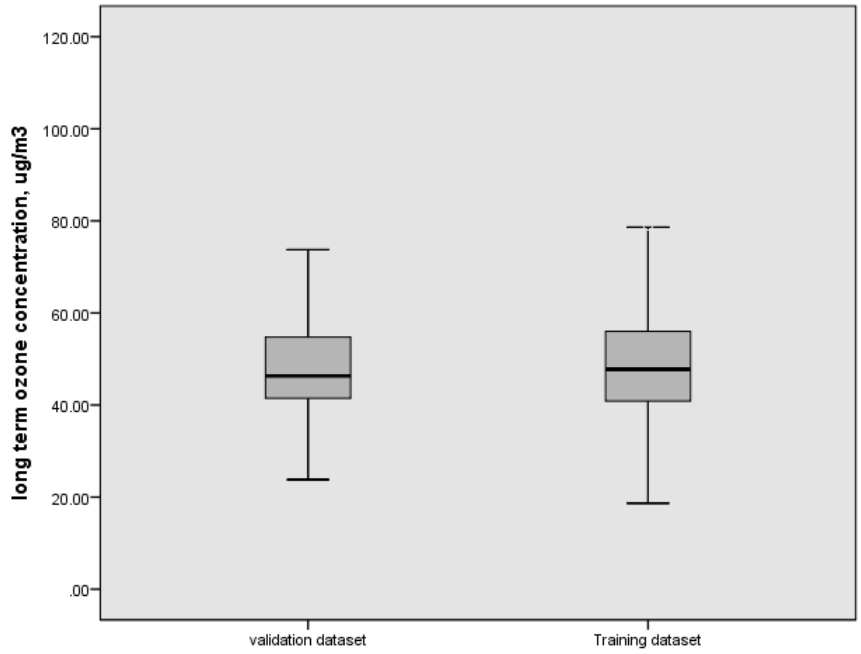


Figure B.1box plot of training and validation datasets

Table B. 4 Descriptive statistics (25, 50, and 75%) for Training and validation datasets.

Statistics	1	2	3	4	5	6	7	8	9	10	11
No. Of Training sites	50	111	114	97	124	48	52	80	96	47	45
25%	57.4	55.2	47.5	39.6	45.3	35.4	29.5	35.7	35.4	38.5	38.4
50%	61.9	59.3	50.8	44.8	48.4	42.1	33.5	40.2	39.6	42.7	45.1
75%	65.7	65.9	54.8	48.7	51.9	47.9	39.4	46.3	43.5	47.0	54.6
No. Of validation sites	17	26	29	19	25	10	13	21	23	16	10
25%	56.4	51.7	44.8	42.4	44.0	37.3	30.6	34.9	38.3	39.4	35.7
50%	59.8	60.4	50.1	47.8	47.1	42.2	34.5	40.1	41.2	42.7	41.9
75%	62.0	65.1	54.4	51.3	51.6	46.4	40.0	46.7	43.0	44.3	59.6

V Nearest sites stratified by country

Table B.5 Correlation between nearest monitoring sites stratified by countries

VAR00001	R	R Square	Std. Error of the Estimate
AT	.325	.106	16.699
BE	.255	.065	5.609
DK	.260	.068	9.709
ES	.005	.000	15.150
FR	.460	.210	11.749
GB	.470	.221	9.186
DE	.314	.099	11.051
IE	.040	.002	12.499
IT	.087	.008	14.551
NL	.467	.218	5.169
PT	.486	.243	8.579

VI Temporal models

Table B.6 Site type2 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.28	25.42	.28	.000
D17	.38	23.73	.10	.000
D2_14	.40	23.43	.02	.000
weekday	.40	23.40	.00	.000
Sunday	.40	23.39	.00	.000
PHF	.41	23.27	.01	.000

Table B.7 Site type3 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.23	29.25	.23	.000
weekday	.23	29.22	.00	.000
Sunday	.23	29.21	.00	.000
D2_14	.43	25.29	.19	.000
PM3	.44	24.98	.02	.000

Table B.8 Site type4 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S3	.31	28.49	.31	.000
weekday	.32	28.39	.01	.000
Sunday	.32	28.37	.00	.000
D16	.45	25.48	.13	.000
D2_14	.47	24.96	.02	.000

Table B.9 Site type5 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.26	27.02	.26	.000
weekday	.26	26.94	.00	.000
Sunday	.26	26.93	.00	.000
D16	.40	24.25	.14	.000
D2_14	.42	23.92	.02	.000

Table B.10 Site type6 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.20	31.62	.20	.000
D15	.46	26.11	.26	.000
D2_14	.48	25.68	.02	.000
weekday	.48	25.61	.00	.000
Sunday	.48	25.60	.00	.000

Table B.11 Site type7 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.20	21.94	.20	.000
weekday	.21	21.75	.01	.000
Sunday,	.22	21.71	.01	.000
D16	.26	21.09	.04	.000
D2_14	.30	20.50	.04	.000

Table B.12 Site type8 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S1	.16	26.12	.16	.000
weekday	.17	26.01	.01	.000
Sunday	.17	25.99	.00	.000
D15	.29	24.08	.12	.000
D2_14	.31	23.69	.02	.000

Table B.13 Site type9 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.22	27.34883	.22	.000
D116	.34	25.07085	.12	.000
D2_14	.36	24.76879	.02	.000
weekday	.37	24.61568	.01	.000
Sunday	.37	24.59696	.00	.000

Table B.14 Site type10 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S2	.27	31.29	.27	.000
Weekday	.28	31.24	.01	.000
Sunday	.28	31.23	.00	.000
D16	.46	26.96	.18	.000
D3_15	.48	26.59	.02	.000

Table B.15 Site type11temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S3	.27	27.85	.27	.000
weekday,	.28	27.76	.01	.000
Sunday	.28	27.75	.00	.000
D16	.46	24.04	.18	.000
D2_14	.51	22.96	.05	.000

TableB.16 Site type12 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S1	.23	21.84	.23	.000
D17	.26	21.39	.03	.000
PHF	.27	21.31	.01	.000
weekday	.27	21.30	.00	.000
Sunday	.27	21.30	.00	.000

Table B.17 Site type13 temporal model summary

Model	R^2	RMSE	R^2 Change	P-value
S3	.43	31.27	.43	.000
D17	.55	27.73	.12	.000
D4_16	.57	27.19	.02	.000
PHF	.58	26.97	.01	.000
weekday	.58	26.91	.00	.000
Sunday	.58	26.89	.00	.000

VIII Separate models for weekday and weekend outputs

Exploring if the tow period weekday and weekend need separate models using data from site group 1, the results show no difference the same function were included in the two period model.

Table B.18 Model Summary for weekday and weekend period in site type1

weekday	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
weekend	S2	.515	.265	.265	25.42125
	D16	.582	.338	.338	24.12523
	D2_14	.586	.344	.344	24.02640
	D4_16	.586	.344	.344	24.02609
weekday	S2	.557	.311	.311	26.29463
	D14	.610	.372	.372	25.10640
	D2_14	.616	.380	.380	24.94150
	D3_15	.616	.380	.380	24.93841

Table B.19 Weekday and weekend Models coefficients for site type 1

period	Model	B	Std. Error	Beta	P-value
weekend	Constant	1.863	0.03		0.000
	S2	21.59	0.042	0.515	0.000
	D14	11.318	0.042	0.27	0.000
	D2_14	3.084	0.042	0.074	0.000
weekday	Constant	-0.783	0.02		0.000
	S2	24.934	0.028	0.557	0.000
	D14	11.042	0.028	0.247	0.000
	D2_14	4.062	0.028	0.091	0.000

Divided the data to night time and day time the data to night (hour= 20-24 and 1-4) and daylight (all other hours) also show no differences.

Table B.20 night and day time Models coefficients for site type 1

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Night time	(Constant)	-.416	.056		-7.380	.000
	S2	21.928	.037	.532	585.914	.000
	PM4	8.693	.170	.155	51.052	.000
	PM2NI2AM8	1.884	.121	.047	15.546	.000
Day-time	(Constant)	-.236	.023		-10.063	.000
	S2	25.193	.029	.556	860.810	.000
	PM4	11.305	.036	.247	311.845	.000
	PM2NI2AM8	3.761	.037	.079	100.359	.000

a. Dependent Variable: O3conc

IX Correlation coefficients matrix

Exploring the correlations between the average environmental characteristics from Table 4.10 for each site types and some of the key elements of temporal models from Table 6.4, and the associated statistics.

Table B.21 Correlation coefficients matrix between environmental factors and key elements of temporal models

Correlation Coefficients Matrix										
Sample size	13	*Critical value (5%*	2.20							
		DistSea	*Urban*	*Agric*	*Non-ag*	*Rural*	*Alt*	*Traffic*	*Topex*	*Radiation*
Syst var (%Tot)	**Pearson Correlation Coefficie**	*0.20*	*0.31*	*0.38*	*-0.48*	*-0.09*	*-0.27*	*0.04*	*-0.66*	*0.28*
	R Standard Error	*0.09*	*0.08*	*0.08*	*0.07*	*0.09*	*0.08*	*0.09*	*0.05*	*0.08*
	t	*0.68*	*1.08*	*1.35*	*-1.83*	*-0.28*	*-0.95*	*0.13*	*-2.94*	*0.97*
	p-value	*0.51*	*0.30*	*0.20*	*0.09*	*0.78*	*0.36*	*0.90*	*0.01*	*0.35*
Seasonal(%S)	**Pearson Correlation Coefficie**	*0.10*	*-0.40*	*-0.32*	*0.57*	*0.31*	*0.59*	*-0.32*	*0.72*	*-0.25*
	R Standard Error	*0.09*	*0.08*	*0.08*	*0.06*	*0.08*	*0.06*	*0.08*	*0.04*	*0.09*
	t	*0.35*	*-1.47*	*-1.12*	*2.29*	*1.08*	*2.44*	*-1.11*	*3.43*	*-0.85*
	p-value	*0.74*	*0.17*	*0.29*	*0.04*	*0.30*	*0.03*	*0.29*	*0.01*	*0.41*
Diurnal(%S)	**Pearson Correlation Coefficie**	*-0.08*	*0.32*	*0.43*	*-0.52*	*-0.21*	*-0.56*	*0.20*	*-0.71*	*0.25*
	R Standard Error	*0.09*	*0.08*	*0.07*	*0.07*	*0.09*	*0.06*	*0.09*	*0.04*	*0.09*
	t	*-0.25*	*1.12*	*1.59*	*-2.00*	*-0.71*	*-2.26*	*0.69*	*-3.37*	*0.85*
	p-value	*0.81*	*0.29*	*0.14*	*0.07*	*0.49*	*0.05*	*0.50*	*0.01*	*0.41*
Seasonal (%Tot)	**Pearson Correlation Coefficie**	*0.23*	*0.05*	*0.08*	*-0.08*	*0.09*	*0.09*	*-0.17*	*-0.11*	*0.01*
	R Standard Error	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*
	t	*0.77*	*0.16*	*0.26*	*-0.28*	*0.29*	*0.29*	*-0.56*	*-0.35*	*0.02*
	p-value	*0.46*	*0.87*	*0.80*	*0.79*	*0.78*	*0.77*	*0.59*	*0.73*	*0.98*
Diurnal (%Tot)	**Pearson Correlation Coefficie**	*0.05*	*0.30*	*0.51*	*-0.52*	*-0.13*	*-0.44*	*0.14*	*-0.77*	*0.37*
	R Standard Error	*0.09*	*0.08*	*0.07*	*0.07*	*0.09*	*0.07*	*0.09*	*0.04*	*0.08*
	t	*0.16*	*1.04*	*1.95*	*-2.02*	*-0.44*	*-1.62*	*0.47*	*-4.02*	*1.32*
	p-value	*0.88*	*0.32*	*0.08*	*0.07*	*0.67*	*0.13*	*0.65*	*0.00*	*0.21*
Sunday increment	**Pearson Correlation Coefficie**	*-0.15*	*0.87*	*-0.47*	*-0.79*	*-0.80*	*-0.74*	*0.78*	*-0.55*	*-0.14*
	R Standard Error	*0.09*	*0.02*	*0.07*	*0.03*	*0.03*	*0.04*	*0.04*	*0.06*	*0.09*
	t	*-0.49*	*5.99*	*-1.77*	*-4.33*	*-4.38*	*-3.67*	*4.12*	*-2.20*	*-0.47*
	p-value	*0.64*	*0.00*	*0.10*	*0.00*	*0.00*	*0.00*	*0.00*	*0.05*	*0.65*
Weekday decrement	**Pearson Correlation Coefficie**	*-0.22*	*0.88*	*-0.40*	*-0.82*	*-0.84*	*-0.75*	*0.79*	*-0.50*	*0.03*
	R Standard Error	*0.09*	*0.02*	*0.08*	*0.03*	*0.03*	*0.04*	*0.03*	*0.07*	*0.09*
	t	*-0.73*	*6.01*	*-1.44*	*-4.78*	*-5.09*	*-3.75*	*4.24*	*-1.94*	*0.10*
	p-value	*0.48*	*0.00*	*0.18*	*0.00*	*0.00*	*0.00*	*0.00*	*0.08*	*0.93*