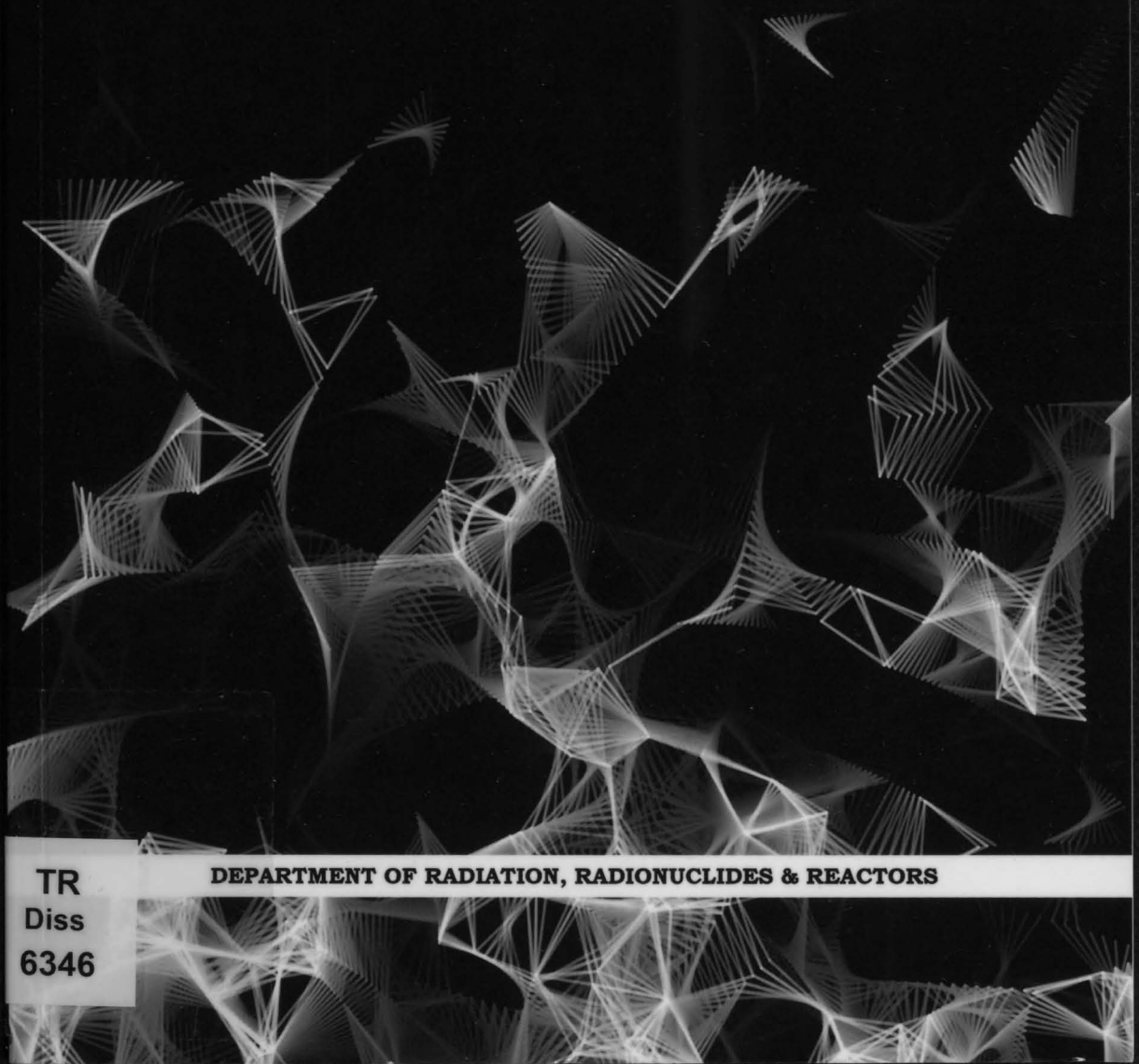
# Application of Atmospheric Biomonitoring to Epidemiology

Issues in Data Quality, Sampling, Aggregation & Confounding

*Susana F.M. Sarmento*





DEPARTMENT OF RADIATION, RADIONUCLIDES & REACTORS

# APPLICATION OF ATMOSPHERIC BIOMONITORING TO EPIDEMIOLOGY

# APPLICATION OF ATMOSPHERIC BIOMONITORING TO EPIDEMIOLOGY

## ISSUES IN DATA QUALITY, SAMPLING, AGGREGATION & CONFOUNDING

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op 25 april 2012 om 10:00 uur
door

**Susana Fernandes de Morais SARMENTO**
MSc Neuroscience, University of Edinburgh
geboren te Porto, Portugal.

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. H. Th. Wolterbeek
Prof. dr. ir. M. C. Freitas

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter
Prof.dr. H.Th.Wolterbeek, Technische Universiteit Delft, promotor
Prof.dr. M.C. Freitas, Instituto Tecnológico e Nuclear, promotor
Prof.dr. E. Steinnes, Norwegian University of Science and Technology
Prof.dr. O. Hänninen, University of Eastern Finland
Prof.dr. E.H. Brück, Technische Universiteit Delft
Prof.dr.ir. M. De Bruin, Technische Universiteit Delft
Dr.ir. P. Bode, Technische Universiteit Delft

Fundação para a Ciência e Tecnologia (Portugal) heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.



ISBN 978-1-61499-048-2

Keywords: biomonitoring, epidemiology, ecological, time-series, sampling, aggregation, confounding

*Published and distributed by IOS Press under the imprint Delft University Press*

*Publisher*
IOS Press
Nieuwe Hemweg 6b
1013 BG Amsterdam
The Netherlands
tel: +31-20-688 3355
fax: +31-20-687 0019
email: info@iospress.nl
www.iospress.nl

LEGAL NOTICE
The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

*To my dear mum, dad and sister*

# Table of Contents

# 1 Introduction

*Partly based on book chapter:*
*"Statistical Approaches in Environmental Epidemiology"*
*by Verburg TG, Sarmento SM & Wolterbeek HTh.*
*In: (S. Lahiri, ed.) Advanced Trace Analysis, pp.1-69 (2010).*
*Narosa Publishing House Pvt. Ltd. New Delhi, India.*

## 1.1 Motivation

It was nearly 150 years ago that William Nylander (1886) noted a decrease in lichen biodiversity in Paris. At about the same time, across the Channel, the word smog entered common parlance and was immortalised in the statement *"...it required no science to see that there was something produced in great cities which was not found in the country and that was smoky fog, or what was known as 'smog'."* (Dr. Henry Antoine de Voeux, 1905; Wikipedia). The first half of the 20th century was sprinkled by reports that began with the words "cold" and "haze" and ended with graphs showing swift increases in deaths and hospital admissions, mostly due to respiratory complaints among the frail. The common denominator in these reports was the combination of thriving industrial activity and transitory winter inversions (e.g.: Meuse Valley, Belgium in 1930; Donora, Pennsylvania in 1948; London, UK I 1952; reviewed by Lipfert, 1993). Only one of these early episodes (London 1942) recorded air pollution levels (smoke and $SO_2$), however, no science indeed was required to prove that air pollution's effects extended well beyond corroded monuments, traffic mayhem and extinction of symbiotic organisms.

From the second half of the 20th century to our days, herculean strides have been taken to regulate and control industrial and vehicular emissions and improve fuels. This resulted in plummeting pollution levels (10-50 times less) in the greatest part of the World that experienced the Industrial Revolution (Greenbaum, 2003; Krewski et al, 2003). The effort appears to have been noted by the lichens in Paris, and indeed those of London (Seaward & Letrouit-Galinou, 1991; Rose & Hawksworth, 1981), and it must also have been noted by human health although no studies appear to have investigated this issue yet (Maynard & Cohen, 2003; Lipfert, 1997, 1998). Nevertheless, even at the comparatively low pollution levels currently enjoyed in most of the USA and Europe, human health effects continue to be attributed to it, by increasingly sophisticated statistical methods and increasingly specialised pollution and health indicators (Brunekreef & Holgate, 2002; Pope & Dockery, 2006).

Environmental exposure to regulated air pollutants such as $PM_{10}$, $SO_2$, and $O_3$ in the general population has been associated with a narrow and borderline risk to human health. Metrics of risk differ across studies, making comparisons somewhat difficult. Nevertheless, excess risks between 1-25% have been reported (Wakefield, 2003; Lipfert, 1997; WHO, 2000). The risk appears to be fairly similar for acute effects (time-series studies) and chronic effects measured in aggregate ecological studies and somewhat larger for chronic effects measured in multi-

level studies of prospective cohorts, but these differences could be due to random variation (Lipfert, 1995).

Despite the relatively small risks, the fact that air pollution exposure is universal and its anthropogenic emission preventable makes it a worthwhile target for public health intervention. When small measures of effect are translated into a measure of impact at the global scale, as for instance the WHO's estimate that half a million excess deaths per year worldwide are due to PM exposure alone (WHO, 2005; Pope & Dockery, 2006), it is difficult not to be concerned.

The precautionary principle not withstanding, some voices have risen to play the devil's advocate. Could the societal costs (e.g.: industrial job losses) of increasing stricter air pollution abatement strategies cost more lives than the ones it saves (Lipfert, 1997; Bluestone & Harrison, 1982)? Others have noted that current air pollution indicators are usually, if not always, the weakest predictor of health effects, shouldn't these stronger risk factors, such as poverty and social inequality, be subjected to abatement strategies first (Hayes, 2003)? Finally, despite the arguable consistency and coherency of effects across studies (Bates, 1992; Pope & Dockery, 2006; WHO, 2000; Rothman, 2002), its smallness makes it prone to being washed away by the myriad of ways in which observational studies can be biased (e.g.: confounding, measurement error) (Wakefield, 2003; Lipfert, 1999). Confounding and measurement error, like effect estimates, could conceivably show consistency across studies (e.g.: urban/rural gradients are global).

On the other hand, the observed effects could be just the tip of the iceberg. This is because health outcomes and air pollution indicators have been rather unspecific, so if more proximal causal pollutants and susceptible populations were found, the effect estimates would be greater and more robust to biases.

Thus, under the auspices of the still unchallenged precautionary principle, the current challenges of air pollution epidemiology include: the identification of the aspects of air pollution that are most adverse, identification of symptoms and populations that are most susceptible, and estimation of dose-response curves capable of recommending limit values for public health protection (HEI, 2002). This thesis shall address the first issue by giving the leading role to chemical elements, including metals, and by asserting, as another major challenge of modern epidemiology, the need for more and better exposure data.

## 1.2 Air pollution

*"But it is naturally toxic"*
*New Yorker cartoon*

The extreme air pollution episodes of the early 20$^{th}$ century rose awareness towards corrosive components such as particles (defined as Smoke, BS, TSP, etc). These air pollution indicators have the longest monitoring history. Particle indicators have suffered important refinements

by progressively focusing on smaller aerodynamic-sized particles because they are better able to enter the lower respiratory system, they tend to have a more toxic composition and because of the development of sampling technology (e.g.: dichotomous samplers) (Kennedy & Hinds, 2002; Greenbaum, 2003; HEI, 2002).

Nowadays, the most widely monitored air pollutants include corrosive and oxidant gases such as $SO_2$, NO, $NO_2$ and $O_3$ and particles measured as $PM_{10}$, $PM_{2.5}$ and increasingly finer sized fractions.

Particles are an important component of the atmosphere for they are the carriers of numerous liquid and solid compounds. Their composition depends on the source, formation process and the physiochemical characteristics of the atmosphere in which they are emitted. The fundamental ingredients of PM are: oxidised or elemental carbon (mostly at the core), metals, organic compounds, biological material, ions and reactive gases (HEI, 2002; Ghio et al, 1999).

It is unclear whether particles *per se* or some restricted aspect of their physiochemical properties are more closely responsible for health effects. Among the most likely culprits are ultrafine particles and metals (NRC, 1998; HEI, 2002; Samet, 2000; Oberdorster & Utell, 2002).

### 1.2.1 The metal hypothesis

"What components (or mixture thereof) of PM are responsible for the observed health effects" has been deemed a priority research question in air pollution epidemiology (NRC, 1998; HEI, 2002; Harrison & Yin, 2000; Dreher, 2000; Schlesinger et al, 2006). Among the usual suspects are metals.

Metals are released into the atmosphere by high-temperature processes such as volcanic activity and combustion of fuel and waste. As the temperature cools down away from the source, the vaporised metal either forms new particles or condenses over existing ones (Avakian et al, 2002). Chemical elements, including metals, may also be released from natural sources such as erosion and sea spray and by anthropogenic activities such as quarrying. In the latter cases elements tend be associated with large particles and tend be to chemically inert (e.g.: silicates).

The metal/elemental content of particles, by mass, has been reported to range from less than 1% in environmental PM to as high as 20% in residual fly ash (Ghio et al, 1999; Roosli, 2001).

Numerous toxicological studies have uncovered biochemical pathways for toxicity of metal-enriched particles (Ghio et al, 2002; Ghio & Devlin, 2001; Dye et al, 1999; Dye et al, 1997; Kodavanti et al, 1998) and metals in environmental PM (Knaappen et al, 2002; Gavett et al, 2003; Gerlofs-Nijland et al, 2009). In addition, numerous toxicological studies have established plausible health outcomes that can be attributed to such exposures (extensively reviewed by WHO IARC Monographs, EPA IRIS Reports and ATSDR Toxicological Profiles, found online). Several chemical elements are accepted carcinogens by inhalation

exposure: As (also by ingestion), Be, Cd, Cr(VI), and Ni. Pb is a probable human carcinogen, whereas Co is a possible human carcinogen (WHO IARC classification).

Perhaps the most compelling evidence in support of what has come to be known as "the metal hypothesis" stems from a collection of studies performed in the Utah Valley that took advantage of a natural experiment occasioned by the temporary closure of a steel mill circa 1982 (reviewed by Ghio, 2004). The collection, which included thorough toxicological (Ghio & Devlin, 2001; Ghio et al, 1999; Kennedy et al, 1998; Dye et al, 2001), and epidemiological investigations (Pope, 1996, 1991, 1989), determined quite unequivocally that the metal composition of PM was responsible for adverse health effects.

Epidemiological investigations in the context of environmental exposures among the general population have been protracted due to the sparse monitoring network. Besides the holistic collection of studies performed in the Utah Valley, and occasional epidemiological studies (Dusseldorp et al, 1995; Lipfert., 1998, 1988, 1980), it was not until recently that researchers began using the limited but growing data on airborne chemical elements. Most studies to date have been of a time-series design and used source apportionment data, rather than the elements themselves, as exposures (Mar et al, 2006; Ito et al, 2006; Thurston et al, 2005; Laden et al, 2000; Claiborn et al, 2002). Owing to the requirement of a wider and denser monitoring network, only a few cross-sectional studies have been performed (Harrison et al, 2004; Lipfert et al, 2006).

It is widely believed that the effects of air pollution are probably stronger for chronic than for acute diseases. In order to assess the chronic health effects associated with long-term exposure to metals the most appropriate design is the cross-sectional. This design, however, demands exposure data over dense and wide geographical scales and for extended periods of time. This is somewhat impractical considering the costs of setting up and maintaining dedicated monitoring networks and the more complex chemical analysis required to determine metal concentrations. AirBase compiles air pollution measurements across the European continent. From 1980-2009, there were between 3000 to 5000 monitoring stations for $PM_{10}$ and the traditional gaseous pollutants. For the same period, there were 500-1000 monitoring stations for Pb, Cd, Ni and As, and at most 51 monitoring stations for other chemical elements such as Hg (EIONET, 2011). Technological advancements, such as solar-powered monitors will certainly decrease the costs of maintaining a sizeable metal monitoring network (Wolterbeek et al, 2010). Another viable, and readily available, approach is biomonitoring.

### 1.2.2 Atmospheric biomonitoring

In the context of "the metal hypothesis" and the need to ensure prolonged and spatially dense exposure monitoring in order to assess chronic health effects, atmospheric biomonitoring stands out as a prominent solution.

Mosses and lichens are believed to be some of the best biomonitors of several atmospheric pollutants, including chemical elements, gases and dioxins. This belief is rooted on two

characteristics of these organisms: 1) they acquire nutrients virtually exclusively from atmospheric deposition (both wet and dry), and 2) they have a simple physiology which makes them relatively passive accumulators (Szczepaniak & Bizuik, 2003; Garty, 2001; Godinho et al, 2008).

The use of mosses and lichens as passive bioaccumulators of atmospheric deposition has a long and distinguished history in Europe, especially in Scandinavia (Ruhling & Tyler, 1968, 1973; see references in Wolterbeek et al, 2010) and has been gaining momentum in other Continents in recent years (e.g.: Morocco: El Khoukhi et al, 2003; Argentina: Pignata et al, 2007; Ghana: Nyarko et al, 2006; China: Lee et al, 2005). Europe-wide moss surveys have been performed periodically since 1977, leading to both geographical and longitudinal descriptive studies of airborne metals (Ruhling, 1994; Steinnes et al, 1994; Garty et al, 2009; Buse et al, 2003; Harmens et al, 2004). In addition, national surveys have been performed in many countries, often also on a periodic basis (e.g.: England: Ellison et al, 1976; Sweden: Ross, 1990; Germany: Markert et al, 1996; Slovenia: Jeran et al, 1996, 2003; Portugal: Freitas et al, 1997, 1999; Figueira et al, 2002; Netherlands: Sloof & Wolterbeek, 1991).

Compared to conventional instrumental monitoring, biomonitoring offers advantages that are difficult to surpass: 1) the ability to perform high-density sampling at virtually any desired spatial and temporal scale, and 2) the ability to measure a wide range of pollutants simultaneously. This is achieved at comparatively low costs and man-power, since biomonitors are energetically self-sustainable, require no maintenance and are not attractive targets for vandalism.

By way of comparison, the 2000/01 European moss survey (Harmens et al, 2004) measured As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, V and Zn at 6380 sampling sites spanning 26 countries, resulting in a sampling density of about 1.3 per 1000 $km^{-2}$ (these figures exclude Russia because the area sampled in Russia was not clear). This is a far cry from the less than 1000, often much less than 100 instrumental monitors measuring metal concentrations throughout Europe (EIONET, 2011).

In addition to their potential for routine air pollution monitoring, mosses and lichens are invaluable tools for "natural experiments" since they are present nearly everywhere at all times (e.g.: Chernobyl accident by Sloof & Wolterbeek, 1992; closure of industrial plant: Rusu et al, 2006; mine exploration: Branquinho et al, 1999).

The chief disadvantages of biomonitoring are tied up with the fact that although lichens and mosses are relatively simple organisms, they are nevertheless far more complex than inanimate hyper-pure filters under known and controlled ventilation protocols. The kinetics of metal retention, ad/absorption and excretion need to be understood, and they may vary depending on the biomonitor's morphology (Godinho et al, 2009a, 2009b). In addition, several factors influence the extent to which biomonitors capture and retain elements, including environmental conditions such as air quality and weather, and characteristics of the biomonitor itself such as its species, age and health (Godinho et al, 2004, 2008; 2011a, 2011b;

Szczepaniak & Bizuik, 2003; Garty, 2001; Wolterbeek, 2001a, 2002; Wolterbeek et al, 2010; Conti & Cecchetti, 2001). To complicate matters all these issues may be element-dependent.

Throughout time, the sampling and analytical methodology has been increasingly standardised and subjected to rigorous quality criteria and assessments, in great part under the auspices of IAEA and UN projects (Ruhling, 1994; Smodis & Bleise, 2002, 2007; Smodis, 2003; Smodis & Parr, 1999; Markert et al, 2003; Harmens, 2010). For instance, for most purposes biomonitors must be sampled at least $50m^{-1}$ away from main roads and at no less than $1m^{-1}$ in height. Furthermore, reference materials of known concentrations have contributed to more reliable measurements (e.g.: IAEA-336 lichen reference material; Heller-Zeisler et al, 1999). Further quality improvements and standardisation of all aspects of the survey's design, from sampling to the analysis is one of the most pressing topics in atmospheric biomonitoring (Wolterbeek et al, 2010; Wolterbeek, 2002), as is systematic and comprehensive research that can lead to knowledgeable decisions (e.g.: Sloof, 1993; Reis, 2001 Marques, 2008; Godinho, 2010).

In what concerns the application of biomonitoring data to epidemiological research, two questions stand out:

1. To what extent do elemental contents in biomonitors reflect atmospheric monitoring, as measured by instrumental methods? Air pollution is expressed in per $m^{-3}$ air by regulations whereas biomonitoring expresses it in per $g^{-1}$ of biomonitor. How can the latter metric be converted into the former?
2. What is the period of time reflected by the elemental contents in biomonitors? What factors influence it and how can concentrations be calibrated against time of exposure and accumulation?

The answer to the last question, on which the first question depends, is bound to vary depending on the chemical element, the biomonitor species and numerous other factors (see references above). Mosses, in particular, were believed to accumulate pollutants over extended periods (2-3 years). Recent research in lichens suggests that accumulation periods may be of just a few months for most elements, but provided that emission and environmental conditions do not change the accumulation period may be longer (Godinho et al, 2008, 2011b; Reis et al, 1999, 2002).

The accumulation period of biomonitors, whether of a few months or years, means that they are suitable for epidemiological studies of chronic health effects, where an annual average has been conventioned as the minimum indicator of long-term exposure. The assumption that air pollution exposure remains constant throughout the years over which health effects are recorded is common in current epidemiological studies using instrumental monitoring data (Lipfert et al, 2000). However, it may be necessary to perform repeated biomonitoring sampling surveys, especially at locations where emissions fluctuate on a long-term basis. In addition, it will be necessary to calibrate the biomonitors' elemental contents against the confounding effects of their physiological status, environmental conditions and local pollution sources such as soil re-suspension (Godinho et al, 2008, 2011a, 2011b; Reis et al, 1999, 2002).

## 1.3 Epidemiological studies

### 1.3.1 Epidemiological studies of regulated air pollutants

The epidemiological designs used to investigate air pollution exposures in the "general" population may be divided into three main types: time-series, multi-level of prospective cohorts and aggregate ecological (Verburg et al, 2010; Rothman, 2002). Time-series are used to study acute health effects as a result of short exposures and short induction periods (one day to one week), whereas the other two designs are used to study chronic health effects as a result of long-term exposures (one year or more) and long induction periods (at least 10 years).

The results of these studies have been reviewed extensively on many occasions (e.g.: Pope & Dockery, 2006; Lipfert, 1993, 1995, 1997; HEI, 2002; Schwartz, 1994a; Thurston et al, 2005; Bates, 1992; Vedal, 1997). The overarching conclusion is that regulated air pollutants such as $PM_{10}$, Sulphates and $O_3$ are associated, and thus are the probable cause, for adverse health effects, despite the relatively low pollution levels found at most study locations. However, the evidence does diverge somewhat over follow-up times and across health outcomes, pollutants and stratification groups (e.g.: gender and age). Effect estimates are difficult to generalise due to the differences in model specifications and exposure reference intervals (Baxter et al, 1997; Lipfert, 1993; Lipfert & Wyzga, 1995b). For the traditional air pollutants (e.g.: TSP, $PM_{10}$, $PM_{2.5}$, $SO_2$, $O_3$, etc), typical excess risks have been in the order of 2-6% for acute effects, 4-8% for chronic effects assessed through aggregate ecological studies, and 8-25% for chronic effects assessed through multi-level studies of prospective cohorts (Lipfert & Wyzga, 1995b; Lipfert, 1995, 1997; Lipfert et al, 2000).

### 1.3.2 Epidemiological studies of elemental air pollutants

In recent years, epidemiological studies have begun to complement traditional air pollutants data such as $PM_{10}$ and $SO_2$ with data on elemental (including metals) pollutants, when the latter are available from monitoring networks. Most epidemiological studies using elemental pollutants have been of a time-series design and used source apportionment rather than the elements themselves as exposures (Mar et al, 2006; Ito et al, 2006; Thurston et al, 2005; Laden et al, 2000; Claiborn et al, 2002). Most of these studies reported the strongest and most robust associations for secondary Sulphate and traffic-related particles, and weak or no association for particles associated with crustal/natural sources.

To the best of our knowledge only two geographical studies have used elemental exposures from monitoring networks. Harrison et al (2004) resorted to historical measurements of As, Ni, Cr and PAH to predict lung cancer mortality in the American Cancer Society (ACS) cohort. They found that estimated effects were within the range attributed to $PM_{2.5}$ in that cohort, i.e. 8-13% excess risk per $10\mu g\ m^{-3}$. Lipfert et al (2006) compiled monitoring data for 15 chemical elements to predict total morality in the Washington Veteran's cohort. They found that apart from peak-$O_3$, Ni and V were the only statistically significant predictors in

single-pollutant models. Interestingly, it was a far less specific pollution indicator, traffic density, that displayed the strongest and most resilient association. The elasticity at the mean for traffic density and peak-$O_3$ was 20% whereas that for Ni and V was 5%.

As with instrumental monitoring of elemental pollution, the use of atmospheric biomonitoring in epidemiology has a short history (Sarmento et al, 2008; Wolterbeek & Verburg, 2004a; Wappelhorst et al, 2000; Cislaghi & Nimis, 1997). The existing studies are all cross-sectional and aggregate ecological, and their results are mostly exploratory because they used correlation coefficients and made no or feeble attempts to control for confounding.

Gailey & Lloyd (1993) were perhaps the first to suggest the use of atmospheric biomonitoring in epidemiology. Cislaghi & Nimis (1997) were the first to use bioindication in epidemiology. In their Nature paper, they correlated the spatial distribution of a lichen diversity index (1991) with the spatial distribution of deaths due to respiratory diseases in 662 municipalities of the Veneto region (Italy) in 1981-88. They found a strikingly strong correlation and neat map overlap for lung cancer mortality in native males <55 years old, but not for any other age, gender, and migration groups or diseases.

Wappelhorst et al (2000) were the first to use atmospheric biomonitoring in epidemiology. They correlated numerous chemical elements determined in mosses (1995 and 1996) with hospital discharges (including deaths) caused by several diseases (1993-97). The area of study was the EuroRegion Neisse, also known as the Black Triangle due to its history of intense industrial activity. However, by the time the moss survey was carried out, the region's metal levels were quite homogeneous and comparable to average levels found in most of Europe. The geographical unit of comparison was unclear, possibly districts, but their number is not mentioned. This study raised several methodological questions, such as: how to convert point exposures into surfaces, whether to analyse aggregated or disaggregated gender-age groups and the importance of having a wide range in exposures in order to obtain statistical significance. They found significant associations between Tl and cardiovascular diseases, and between Ce, Fe, Ga and Ge and respiratory diseases.

Wolterbeek & Verburg (2004) was the second study to use atmospheric biomonitoring in epidemiology and the first to use source apportionment results in addition to the individual chemical elements. They correlated chemical elements measured in mosses (1995) and their emission factors calculated by Monte Carlo Target Transform Factor Analysis (MCTTFA) with mortality due to numerous causes, averaged over 1993-95. The unit of analysis were 10-11 provinces in the Netherlands. Extraneous variables (e.g.: address density) were used to confirm the emission factors' identification. This study raised the question of how to interpret divergent associations between aggregated and disaggregated exposures. In particular, they observed that a disease could be associated with an emission factor, but not with its main component elements, or conversely, that a disease could be associated with a chemical element but not with its main emission factor. They detected a negative association between Se and mortality due to neoplasms, circulatory and digestive diseases and a positive association between Br, Cl and Na and mortality due to genito-urinary diseases. "Natural"

emission factors, attributed to soil and lichen physiology, were not significantly correlated with any of the health outcomes.

Sarmento et al (2008) correlated 39 chemical elements measured in lichens in 1993 and deaths due to cancer. The unit of analysis was 25 NUTS-III regions in Portugal. An exploratory attempt was made to control for confounding variables, through forward stepwise F-change selection. Volatile combustion-related elements (Br, I, Ni, Pb, S, Sb and V) were found to be significantly associated with cancer deaths.

## 1.4 Ecological studies

Apart from perhaps laser beams and cold nuclear fusion, few disciplines have deserved such numerous cycles of credulity and scepticism as ecological studies. Their vast potential is thwarted by comparatively trifling difficulties.

Most of the research in the epidemiology of environmental exposures to air pollution is based on ecological analyses, mostly because air pollution cannot be measured for each individual on a large scale but also because they are cheaper and are more widely representative of the human population, including its susceptible subgroups.

The word ecological is used here to refer to the unit of analysis. In time-series studies, the unit of analysis are people grouped in time, usually days. Because on each day, the population is the same and most of its characteristics do not change greatly, the population acts as its own control. As a result, time-series studies are unlikely to be confounded by factors such as lifestyle and socioeconomic factors. Instead, confounding bias may arise from risk factors that vary on a short time-scale such as temperature, and factors that vary on seasonal time-scales such as influenza epidemics. In cross-sectional studies, on the other hand, the unit of analysis are people grouped in space, usually geo-political regions. Because each region contains different populations, which may differ with respect to numerous characteristics such as age-structure, lifestyle and socioeconomic factors, the sources of confounding bias are much more varied and complex than in time-series studies (Lipfert, 1997).

In this section the focus shall be on just prospective cohort and cross-sectional studies.

In the early days, cross-sectional aggregate ecological designs were the norm (Lipfert & Morris, 2002; Lipfert, 1995). As the dismay over the inferential problems presented by these studies grew, the spotlight turned to multi-level studies of prospective cohorts. Cohort studies are very expensive and lengthy and so there are very few of them (about five have been used in air pollution epidemiology). These cohort studies are still ecological, since air pollution is not measured at the individual-level, however these studies have the fundamental advantage of being able to measure individual-level confounders and to control them at the individual-level of analysis (e.g.: Dockery et al, 1993; Pope et al, 1995; Abbey et al, 1999; Lipfert et al, 2006; Jerrett et al, 2003; Krewski et al, 2005).

Ecological studies are useful. There are numerous success stories for the role of ecological studies as hypotheses generators of otherwise unsuspected relationships (e.g.: snuff dipping

and oral cancer by Blot & Fraumeni, 1977). They are also well suited for quantifying measures of impact, required by public health agencies, and to study rare health outcomes. Finally, albeit less consensual, they can add to the "total body of evidence" in the case of already well-established risk factors.

Ecological studies have advantages over studies of individuals, although the advantages are more due to a collateral effect of earthly convenience than to the pursuit of blue-sky science. First, they use routinely collected data and so are less expensive, less labour demanding and faster to complete. Second, their use of spatial and temporal aggregate variables has four advantages: 1) greater ranges in exposures, 2) averaging of variables muffles noise and measurement error, 3) evaluation of variables that are difficult to measure at the individual level (e.g.: air pollution, latitude) or that have no individual-level counterpart (e.g.: income disparity, population density), and 4) study population is more representative of the true population. The first and second advantages imply that (all else being equal) ecological studies may be statistically more powerful than individual ones. Finally, an important advantage of ecological studies is that they can avert confounders (e.g.: gender) (Wakefield, 2003; Greenland & Morgenstern, 1989). Several researchers have highlighted the benefits of ecological analysis, in spite of its pitfalls (Wakefield & Salway, 2001; Wakefield, 2008; Piantadosi, 1994; Cohen, 1994).

Ecological studies are, however, most infamous for their disadvantages. Some ecological data is more prone to systematic measurement error and covariates tend to be more highly collinear at the aggregate level. The most important disadvantage, however, are the complex biases stemming from the mismatch between the level of analysis (groups of individuals) and the desired level of inference (individuals) (Greenland & Robins, 1994a, 1994b).

Most research questions in environmental epidemiology refer to biological processes such as disease, which by definition, act at the level of individuals. However, the analysis is performed on aggregates of individuals formed by spatial proximity (cities, metropolitan areas or provinces) and/or time proximity (days in the case of time-series). It has long been known that associations found at the ecological level do not necessarily apply to the individual-level (and vice-versa), as demonstrated by the classic works of Émile Durkheim and William S. Robinson (Durkheim, 1979; Robinson, 2009). This flaw in reasoning is known as a fallacy, in this context, the ecological fallacy. The quantitative version of this logical problem is called the ecological bias, which is defined as the difference between the "true" estimate and that given by aggregate data (Greenland & Morgenstern, 1989; Morgenstern, 2008; Salway, 2003; Firebaugh, 1978; Webster, 2007). The realisation of this bias, likely compounded by the cultural zeitgeist to the west of the Iron Curtain (Subramanian et al, 2009), led to the downfall of studies that used communities instead of individuals. However, researchers persevered on understanding exactly why and in what conditions ecological bias arises. They found that the "true" estimate is often not adequately estimated from individual-level data either, and in fact multi-level data will often be required. Accordingly this became known as the individualistic fallacy (Subramanian et al, 2009; Firebaugh, 1978; Webster, 2007). The ecological and individualistic fallacies are known collectively as the cross-level fallacy.

### 1.4.1 Ecological biases

*"The technology involved in making something properly invisible is so mind-bogglingly complex that 999,999,999 times out of a billion it's simpler just to take the thing away and do without it..... The "Somebody Else's Problem field" is much simpler, more effective, and can be run for over a hundred years on a single torch battery."*
*Douglas Adams*

Since the effect of some exposure is the quantity of interest in epidemiology, the word bias refers to it. Ecological bias means that the effect of the exposure of interest has been estimated from an ecological analysis and thus may be different from the "true" estimate. Much of the discussion presented here only applies to linear models where all variables are continuous (Salway, 2003; Wakefield, 2003; Webster, 2007; Glynn et al, 2008).

The current interpretation of ecological bias is that it is an umbrella term for several biases, most of which are also present in individual-level studies, especially if the latter recruit individuals from a wide range of locations. As a result, individual-level studies have gradually been replaced by multi-level studies, whereas multi-level studies have gradually been improved by incorporating other ecological variables in addition to the exposure of interest (Jerrett, et al, 2003; Lipfert, 1997; Lipfert et al, 2000; Oakes, 2009). Ecological studies, however, can not easily become multi-level, because it is often impossible to obtain individual-level data. Not only is it nearly impossible to correct for ecological bias without individual-level data, it is often impossible to even detect it. Furthermore, biases are far larger and unpredictable in ecological studies (Salway, 2003; Greenland & Morgenstern, 1989; Webster, 2007).

Numerous studies have been performed to establish the conditions under which the ecological effect estimate is equal to the "true" estimate (the latter may be individual-level or multi-level). The most important requirements include: the use of unstandardised regression coefficients instead of standardised ones (Firebaugh, 1978; Subramanian et al, 2009); and the use of ecological units where the exposure is homogeneous (Greenland & Morgenstern, 1989; Salway, 2003; Webster, 2007). If all individuals in the groups are equal with respect to the exposure, an ecological study boils down to a study of individuals, and will suffer from biases typical of individual-level studies.

Ecological bias has been decomposed into several biases, including (reviewed in Salway, 2003): measurement error (Zidek et al, 1996; Zeger et al, 2000; Lipfert, 1999; Prentice & Sheppard, 1995; Morgenstern, 1982; Brenner et al, 1992; Greenland & Brenner, 1993; Greenland, 1980), mutual standardisation bias (Rosenbaum & Rubin, 984), misspecification of variables (Greenland & Robins, 1994a). However, the most important biases are confounding and the related effect modification, of which there are two versions: between-area confounders/modifiers and confounding/modifiers by group.

#### 1.4.1.1 Confounding and effect modification

Confounding bias (and effect modification) affects ecological studies more severely than individual-level studies, not only because effect estimates are more greatly biased compared to the true estimate, but also because the bias will tend to be upwards. This occurs because ecological studies are: 1) unable to access micro-data and 2) lose information to aggregation (Webster, 2007; Reynolds, 1998; Salway, 2003).

Although it does seem likely that the severity of confounding bias in ecological studies may have been exaggerated by a successive string of contrived numerical and conceptual examples, that are unlikely to arise in practice (Greenland & Robins, 1994a; Greenland & Morgenstern, 1989; Lipfert, 1997; Webster, 2007), it has been recommended that unless the exposure of interest has a large expected effect (Relative Risk>1.4 has been suggested), one should simply refrain from carrying out an ecological study (Wakefield, 2003).

Because the level of inference is individuals, the ecological model must be specified with the same confounders and effect modifiers that would be considered if the study had been individual-level (Salway, 2003). The individual-level confounders/modifiers are then termed between-area confounders/modifiers. Many individual-level risk factors, such as lifestyle and physiology, are not recorded as keenly as income and wealth by statistical agencies. Therefore, ecological studies often do not have data on all the required confounders. Individual-level studies, on the other hand, can obtain whatever data they desire through questionnaires or interviews to cohort participants.

Even if ecological data on all the required individual-level confounders could be obtained, what is the best way to specify the variables at the group level? For instance, smoking at the individual-level can be fully characterised by recording cigarettes x years for each cohort participant, perhaps with an additional term for age at which smoking began. At the group-level this variable would usually be specified as either % smokers or mean cigarettes smoked; neither of these statistics can fully characterise the smoking distribution within groups; in combination, however, they can (Greenland & Robins, 1994a). Often statistical agencies will not collect/provide sufficient data on a variable that enables its complete specification at the aggregate level.

Finally, even if ecological data on all confounders required by the individual-level model could be obtained and even if they could all be adequately specified, confounding bias would still not be completely controlled for. This is because aggregation changes the "meaning" of the confounders and their relationship with the exposure (Firebaugh, 1978). As a result, variables that cause confounding at the individual-level may not cause confounding at the ecological variable (e.g.: gender), i.e. individual-level confounders need not be between-area confounders. Furthermore, and more importantly, variables that do not cause confounding at the individual level may cause confounding at the ecological level. These are known as confounders/modifiers by group. For instance, groups are usually formed by place of residence. Suppose that health facilities happened to be correlated with the exposure across groups, so that people living in areas with good health facilities would be at a lower baseline

risk for diseases (known as confounding by group) or the exposure of effect would be lower (known as effect modification by group), compared to people residing in areas where health facilities are worse. In this example, the confounder (health facilities) acts at the group level. The realisation that group-level confounders can also confound individual-level studies, led to the emergence of the multi-level design. In ecological studies, however, it is impossible to detect and control confounding and effect modification by group with ecological data alone, since it would require the estimation of more parameters than the number of observations. Because the grouping process can lead to the emergence of additional confounders/modifiers, and because they cannot be detected, it is recommended that ecological studies include as many potential confounders as possible, in essence anything that distinguishes the groups being compared or equivalently that might be correlated with the grouping process and could account for differences in baseline risk or in the effect of the exposure of interest (Morgenstern, 2008; Salway, 2003; Firebaugh, 1978; Sheppard, 2003).

In reality, confounding and effect modification by group are simply a matter of specifying the adequate variables in the model; if the group-level confounders/modifiers were known and ecological data for them could be obtained, the bias would disappear. The problem is that, with ecological data only, there is no way of knowing whether this is indeed the case. However, the likelihood and magnitude of the bias can be reduced by changing the way the groups are formed. If the groups are formed so that the between-groups variability in the exposure of interest is maximised, and if possible the between-groups variability in the confounders is minimised, confounding and effect modification by group may be prevented in an ecological study (but not in an individual-level study). Currently, the limiting factor in changing the grouping process for ecological analyses are the data providers, since they only provide health and confounder data for geo-politically-defined groups. Other techniques to detect and correct for ecological biases rely on data from samples of individuals or simulations (Salway, 2003; Wakefield & Salway, 2001; Glynn et al, 2008; Wakefield & Haneuse, 2008; Best et al, 2001; Jackson et al, 2006).

#### 1.4.1.2 Methods to detect and control confounding

Confounders and effect modifiers may be avoided to some extent by selecting a particular study population (e.g.: Abbey et al, 1999 used a cohort formed by 7$^{th}$ Day Adventists who are known for not smoking or drinking) or by stratification (e.g.: gender-age groups). However, the detection and control of most confounders can only be made through mathematical modelling.

The methods to detect confounders are the same as to correct it. There are essentially three automated criteria: the Change in Estimate (CE), F-change and the Akaike and Bayesian Information criteria (AIC & BIC).

The CE criterion is considered the most appropriate method for confounding control. It is a stepwise method that usually starts from the saturated model and removes variables one by one. The removal of a confounder from the model is dictated by the change it causes on the effect estimate of the exposure of interest. It has been conventioned that a confounder that

causes a change greater than 10-25% qualifies as a confounder and should be left in the model; otherwise it is removed from the model (Jorgensen et al, 2007; Jerrett et al, 2003). The CE criterion tends to yield smaller models with higher levels of collinearity than statistical-based method (F-change, AIC/BIC).

The F-change criterion evaluates variables in terms of their predictive value (i.e. ability to contribute to model fit), rather than their confounding effect. Studies comparing the F-change and CE criteria have suggested that if the statistical significance for F-change is set at a high value, usually p-value>.20 or .25, the effect estimates will be reasonably similar to those provided by the CE method (i.e. unconfounded) with the added advantage of increased precision and model fit (Mickey & Greenland, 1989; Maldonaldo & Greenland, 1993; Greenland, 1989).

The AIC and BIC are identical in spirit to the F-change criterion in that they evaluate the predictive power of cofounders, rather than their confounding effect. They are particularly useful when comparing models that are not nested.

In general, however, a strong case has been made against variable selection whatever the selection criteria used (Chatfield, 1995; Breiman, 1992; Jorgensen et al, 2007; Chen, 1999). The uncertainties incurred from multiple testing, the measurement error in variables (Zidek et al, 1996; Zeger et al, 2000) and collinearity (Cohen et al, 2003), can lead to erroneous and volatile variable selection. It is best to choose a model a priori based on substantive reasoning or at most use causal diagrams to filter out variables (Pearl, 1995, 1998, 2000).

## 1.5 Thesis outline

*"Not everything that counts can be counted and*
*not everything that can be counted counts."*
*Sign hanging in Einstein's office at Princeton*

The setting for this thesis is the concern that exposure to environmental (outdoor) air pollution could be responsible for adverse health effects in the human population. The populations used in this thesis are either half of the inhabitants of Portugal in 1994-2004 or the inhabitants of Lisbon in 1999-2004. Air pollution indicators consisted of chemical elements measured through biomonitoring with lichens in 1993, for the first population, and regulated air pollutants ($PM_{10}$, $SO_2$, CO, NO, $NO_2$ and $O_3$) measured by the official instrumental monitoring network. All analyses were based on models that were linear in both variables and parameters.

The primary aim of this thesis, however, is not the estimation of effect measures of air pollution, but rather to explore how uncertainties regarding the data, variable selection, aggregation and confounding might weigh on those estimates. The primary impact of this study is to provide some evidence-based advice on how to collect and analyse data in future studies of air pollution epidemiology.

Chapter 2 is concerned with the effect of aggregation of variables in a time-series design. The usual daily level of analysis in time-series studies is prone to noise and outliers. Aggregation of the variables into longer periods of one week was investigated as a means to obtain more robust effect estimates. This chapter is set in the city of Lisbon in the years 1999-2004 where a time-series design is implemented to investigate the relationship between levels of traditional air pollutants (PM10, SO2, NO, NO2, CO and O3) and hospital admissions due to respiratory- and cardiovascular-related diseases.

Chapter 3 presents recommendations for future sampling surveys of air pollution with regards to the number of samples and the sampling grid that should be used under a variety of population distributions, to obtain estimates for the mean, survey-variance and local-variance under chosen margins of error and statistical significance levels.

Chapter 4 is concerned with the robustness of regression coefficients to: 1) data uncertainties and 2) model specifications and 3) model specification uncertainty.

This chapter is set in about 125 municipalities of Continental Portugal over the years 1994-2004 where an ecological cross-sectional design is used to investigate the relationships between chemical elements measured through lichen biomonitoring and hospital admissions due cardiovascular-related diseases.

Chapter 5 uses the same setup and data as Chapter 5 but explores the presence and impact of negative confounding situations on effect estimates, errors and model fit.

Chapter 6 provides a summary and discussion of the results and some suggestions for future research.

This thesis is based on an ecological aggregate framework, where inferences about processes that occur at the individual level are based on group-level data. Thus most predictor variables, including the exposure of interest, are used as surrogates for individual-level variables. Most chapters use a cross-sectional design, except for Chapter 2 which uses a time-series design. The data and level of analysis refer to groups of individuals formed over space (municipalities) or over time (days). The health outcomes are counts of cases of disease, standardised by the resident population in the case of the cross-sectional design. Only non-infectious diseases were considered, so that individual health events may be considered independent. Models included only continuous untransformed variables with no interaction terms. Only single-pollutant models were considered. Estimation was performed with Ordinary Least Squares linear regression.

# 2 Robustness of different regression modelling strategies in epidemiology: a time-series analysis of hospital admissions and air pollutants in Lisboa (1994-2004)

*Based on article of same title:*
*Sarmento SM., Verburg TG, Almeida SM, FreitasMC & Wolterbeek HTh.*
*Environmetrics (2011) 22, 86-97.*

## 2.1 Abstract

Studies of the acute health effects of air pollution have used exposure windows of different spans and related them to single-day responses. Little is known about whether an increased response window span might be a viable alternative to single-day responses. Our aim is to compare a new model specification where both the exposure and response variables are represented as 7 day moving averages (CMA&CMA model) with the most widely used model specifications in the literature, where the response variable is usually a single-day, in terms of coefficients and their precision and robustness. To this end, daily series of 12 emergency-related hospital admissions and 6 air pollutants spanning 5.5 years in Lisbon were analysed through single-pollutant linear regression and, when necessary M-estimation. With our data, the CMA&CMA model yields coefficients that are very close to models where only the exposure variable is specified as a moving average whether the latter are estimated by OLS or robust M-estimation. In addition, the CMA&CMA model leads to more precise and robust estimates than other model specifications. The new model specification is a straightforward tool for adjusting weekend effects and errors. It is also analogous to robust estimation, with the added advantages of being sensitive to extreme values that are clustered in time, and leading to more precise and robust estimates without loss of high-frequency information. One drawback is the induction of autocorrelation in the residuals.

## 2.2 Introduction

Environmental protection agencies recommend averaging times for each air pollutant (AP) on the basis of their usefulness for specific purposes (e.g. acute or chronic human health protection or vegetation protection) and on considerations of the time-scales at which APs fluctuate (WHO, 2005). For the purpose of human health protection, the choice of the averaging times for APs appears to be trapped in a circular reasoning because most epidemiological and toxicological studies tend to use the recommended averaging times whereas the recommended averaging times are based on epidemiological and toxicological studies. The regulatory agencies themselves acknowledge that some degree of subjectivity underlies the setting of this recommendation (WHO, 2005). Historically, the recommended averaging time for acute human health effects has been 24 h, which is the minimum time-unit for which clinical health data are routinely available. Recently, however, epidemiological studies have been using exposures longer than 1 day, either in the form of moving averages

(MAs) or distributed lags (DLMs) and have consistently found that effects of APs increase as the time-span of the exposure window is increased (e.g. Goodman et al, 2004 and references therein). Such studies, though using exposures of several days or even weeks, always specify the health response variable as a single-day (known exceptions to use are Roberts, 2005; Schwartz, 2000a). In this context, we present a new (except for Schwartz, 2000a) model specification (CMA&CMA) where both the exposure and response variables are 7 day centred MAs. In this article we first present the a priori motivation for choosing the CMA&CMA model specification and then we compare it with the most widely used model specifications in the literature, where the response variable is always specified as a single-day, whereas the exposure variable is specified as either single-days at different lags, MAs or DLMs. Because our aim is to compare model specifications, which differ solely in the way the exposure and response variables are specified in time, our modelling strategies diverge somewhat from those commonly found in studies aimed at causal inference or prediction.

## 2.3 Methods

### 2.3.1 Data description

All data spanned the period between the 1st January 1999 and 30th June 2004.

Daily counts of hospital admissions (HAs) in 7 public hospitals in Lisbon were kindly provided by the Administração Central do Sistema de Saúde in Portugal and were aggregated into the following diagnostic categories (ICD9-CM): respiratory (460-519), circulatory (390-459) and cardiac diseases (390-429) and into the age-groups: <15, 15-64, >64 and total; yielding a total of 12 HAs categories.

Hourly concentrations of six APs: $PM_{10}$, $SO_2$, NO, $NO_2$, CO and $O_3$ were obtained from the National Environmental Institute. For each monitoring station, daily concentrations were obtained by calculating the 24 h mean for $PM_{10}$ and $SO_2$ and the daily 1 h maximum for the other APs. Spatial averaging across Lisbon was performed over the three central-site monitors that measured all APs throughout the entire study period: Avenida da Liberdade, Entrecampos and Olivais. These calculations were performed according to the recommendation of WHO (1999).

Daily mean temperature and relative humidity were obtained from the National Institute of Meteorology.

### 2.3.2 Model specifications

The aim of this paper is to compare the statistical output of the five model specifications described below, by running each model specification on the 72 HA-AP relationships available (12HAx6APs). The five model specifications differ in terms of the way the exposure and response variables are specified in time.

1. CMA&CMA: HAs and APs are both specified as 7 day centred moving averages (CMAs). Model includes only two terms: an AP and a constant, since weekday-

related fluctuations are smoothed by the 7 days window of the CMA. This model specification is the reference model against which all other model specifications below are compared with, because this is the least known model specification.

2. O&CMA: HAs are specified as single-day values, whereas the APs are specified as CMAs. Model includes three terms: a constant, a workday dummy and an AP.
3. O&DLM: HAs and APs are both specified as single-day values. Model includes a constant, a workday dummy and seven terms for the AP lagged from 0 to 6 days. Estimation was performed without constraining the coefficients since we were only interested in the net-slope across all lags.
4. Single-day lags: HAs and APs are both specified as single-day values. Model includes a constant, a workday dummy and one term for the AP at one of 0 to 6 lags. Therefore, for this model specification and for each AP, seven single-day lag models were computed, one for each lagged value of the AP.
5. Best-lag: single-day lag model that, for a given HA-AP relationship, yielded the highest slope among the seven single-day lag models described in item 4. The use of the highest slope as the criterion for choosing the 'best' lag is a usual procedure in the literature (Lumley and Sheppard, 2000).

Because CMA&CMA and O&CMA models include backward lags, additional model specifications were also assessed, namely: FMA&PMA and O&PMA models. The results of these models are mentioned when relevant but their formal results are not presented for the sake of brevity. We preferred to present the results of models using CMA variables because CMA does not induce a shift in the series and, as a result, they are more strongly correlated with the original variables than FMA or PMA variables.

### 2.3.3 Data manipulation

The 7 day centred moving average (CMA) database was created from the original (O) database by replacing each daily observation by the mean over k=7 days (window) and attributing this mean to the (k+1)/2 day. The 7 day prior and forward moving average (PMA and FMA) were calculated in the same way but the mean was attributed to either the $7^{th}$ (PMA) or $1^{st}$ day (FMA) of the window. CMA, PMA and FMA were only calculated when there were no missing values within their windows.

Because our aim is to compare model specifications, it is essential that slopes are comparable between them. To achieve this we normalised all variables in all databases (original, CMA, PMA and FMA) to mean one. Slopes obtained from such normalised variables correspond to elasticities and may be converted into their original slopes by: Elasticity= $\beta \bar{X}/\bar{Y}$ (e.g. Lipfert, 1993).

Although our aim is not causal inference, we felt it important to adjust for any eventual seasonal patterns. We opted for month stratification because it requires few assumptions, adjusts for confounding, effect modification and non-linearities simultaneously and yields simple estimates that are easy to compare. Month stratification does not, however, adjust for potential inter-annual trends. Nevertheless, the method and extent of long-wave adjustment are believed not to be important for the purposes of this article as it will affect all model

specifications being compared in a similar manner. Furthermore, because the results of the comparison between model specifications are identical for all month strata, we present results only for the January stratum (January1999, January2000,..., January2004). Since month stratification implicitly controls for weather effects, and temperature and humidity show very low variability in the January stratum (Table 2.1), these two variables were not considered in any of the analyses.

Table 2.1 Descriptive statistics of the daily levels of air pollutants (μg $m^{-3}$), temperature (°C), relative humidity (%), and hospital admissions counts in the January stratum of the original dataset.

| | | N | Sum | Mean | Med | Min | Max | SD -O[a] | SD-CMA[a] | K-S[b] |
|---|---|---|---|---|---|---|---|---|---|---|
| $PM_{10}$ | | 186 | | 49.27 | 43.79 | 11.05 | 152.24 | 25.73 | 15.46 | 0.111 |
| $SO_2$ | | 180 | | 11.07 | 5.67 | 0.04 | 124.47 | 15.18 | 10.82 | 0.230 |
| NO | | 186 | | 233.64 | 174.13 | 8.67 | 799.57 | 194.11 | 101.08 | 0.167 |
| $NO_2$ | | 186 | | 95.11 | 87.18 | 24.67 | 270.67 | 46.21 | 32.31 | 0.139 |
| CO | | 186 | | 2910.09 | 1980.33 | 363.00 | 11102.67 | 2498.30 | 1669.07 | 0.180 |
| $O_3$ | | 153 | | 40.67 | 40.00 | 4.00 | 77.50 | 17.67 | 12.78 | 0.064* |
| Temp | | 186 | | 11.07 | 11.25 | 4.50 | 16.75 | 2.53 | 1.86 | 0.053* |
| Hum | | 186 | | 79.81 | 81.13 | 48.75 | 98.00 | 10.25 | 6.77 | 0.089 |
| Respiratory | <15 | 186 | 382 | 2.05 | 2 | 0 | 8 | 1.78 | 0.86 | 0.185 |
| | 15-64 | 186 | 1954 | 10.51 | 10 | 3 | 23 | 3.98 | 1.63 | 0.104 |
| | >64 | 186 | 1122 | 6.03 | 5 | 1 | 15 | 2.67 | 1.15 | 0.138 |
| | Total | 186 | 3458 | 18.59 | 18 | 5 | 34 | 5.78 | 3.01 | 0.076 |
| Circulatory | <15 | 186 | 2815 | 15.13 | 15 | 5 | 31 | 4.69 | 2.79 | 0.086 |
| | 15-64 | 186 | 1674 | 9.00 | 9 | 2 | 21 | 3.54 | 1.56 | 0.098 |
| | >64 | 186 | 1082 | 5.82 | 6 | 0 | 17 | 2.79 | 1.21 | 0.100 |
| | Total | 186 | 5571 | 29.95 | 29 | 13 | 49 | 7.29 | 3.44 | 0.068* |
| Cardiac | <15 | 186 | 2629 | 14.13 | 14 | 4 | 27 | 4.47 | 2.68 | 0.084 |
| | 15-64 | 186 | 905 | 4.87 | 5 | 0 | 14 | 2.58 | 1.07 | 0.130 |
| | >64 | 186 | 866 | 4.66 | 4 | 0 | 14 | 2.43 | 1.07 | 0.125 |
| | Total | 186 | 4400 | 23.66 | 23 | 9 | 40 | 6.21 | 2.92 | 0.065* |

[a]Standard deviation (SD) of the original (O) dataset and of the 7 days centered moving average dataset (CMA). [b]Kolgomorov-Smirnov (K-S) statistic for the original dataset where * denotes that the null hypothesis of normality cannot be rejected at the 5% significance level.

Adjustment for weekday effects was made by using a single dummy variable distinguishing weekends from workdays, which captures the major weekly fluctuation. To reiterate, because our aim is to compare model specifications rather than perform causal inference, it is irrelevant whether we use one dummy or the traditional six dummy variables for each weekday type, as long as all model specifications are treated in the same way the result of the comparisons remains unchanged (results not shown).

### 2.3.4 Statistical analyses and software

Linear (OLS) regression was used to estimate the 72 HA-AP relationships across the five model specifications. While most studies have used Poisson or Negative Binomial regression to constrain predictions to positive values, our aim is neither causal nor predictive research, thus this advantage is irrelevant (e.g. Chapter 10 Rothman, 2002). Furthermore, slopes from linear regression may be approximated to Poisson or Negative Binomial slopes times the mean of the dependent variable (e.g. page 89 Cameron and Trivedi 1998). As a check, we compared the slopes (using the approximation) and t-values of the slopes obtained from linear regression and Poisson regression with a paired-sample t-test: we found no statistically significant (1%) differences between these two types of regression for O&CMA and CMA&CMA models.

In a final comparison we performed M-estimation using Huber's and Tukey's Biweight weighing schemes. M-estimation was used because it has become a common procedure in the literature and one that has an impact on the value of all statistics including the slopes.

The statistical significance level for evaluation of each HA-AP relationship within each model specification was set at 1%, owing to the large number of significance tests performed; for all other tests a 5% significance level was used.

Analyses were performed in: Excel 2003, MatLab 7.0.1 and R 2.6.2.

## 2.4 Results

Descriptive statistics are presented in Table 2.1.

### 2.4.1 Reasoning for the a priori choice of the CMA&CMA model

In this section we present the two major theoretical reasons that motivated us to evaluate the CMA&CMA model specification, namely: the ability of MAs to smooth potential influential values and noise in the data and the possibility that ecological studies may lack the temporal resolution to link responses and exposures on such a fine scale as single days. We end the section with a description of some anticipated disadvantages of the CMA&CMA model.

#### 2.4.1.1 Noise and errors

The epidemiologist rarely participates in the sampling or recording of data and often has no access to retrospective information on the sources of inaccuracies and imprecision. For these reasons, major investments should be made on inspecting data quality and potentially adapt the analyses to these inspections. One such adaptation is the literature's regular use of M-estimation, which gives less weight to extreme values in the dependent variable (e.g. Samoli et al, 2001; Schwartz, 2000b).

In the case of APs, measurements require good quality assurance practices, as they are vulnerable to a wide range of technical problems and meteorological influences (EEA, 1998). Although daily AP concentrations are averaged over hourly measurements and monitors,

which smoothes any eventual unusual observations, our data presents several instances of values that may be considered suspicious, as illustrated in Figure 2.1.

HAs, on the other hand, are known to be particularly vulnerable to errors during data recording, but they are also subject to other more complex interferences. It must be borne in mind that the diagnostics are relatively unspecific to APs and even to acute AP exposures. Even after seasonal adjustment, a fraction of daily HAs may be due to APs acting at other time-scales (sub-daily to chronic) and another fraction may be due to other health determinants altogether (assuming a single causal factor). Errors, mixing of time-scales and health determinants, allied to the rare count nature of HAs which is summed rather than averaged, over time and hospitals, may contribute to the deterioration of data quality. Our HAs data, however, appear to be rather devoid of outliers and though they do present autocorrelation at frequencies $\leq$7 days, part of their variability may be noise (Figure 2.2).

Since extreme values, which can be either much higher or much lower than the mean, as well as noise could have dubious impact on the results, the advantages of the CMA&CMA model specification are that it is able to smooth both these effects and in both the exposure and response variables. Furthermore, because each data point becomes an average of seven observations, the error associated with each observation is smaller. Consequently, one would expect an increased accuracy and stability of estimates with the CMA&CMA model specification compared to other model specifications. Another important property of MAs is that it smoothes extreme values that are isolated to a greater extent than extreme values that are clustered in time, as illustrated in Figure 2.1 and 2.2. This is desirable because, in the absence of further information, the first situation is more likely to reflect an error than a 'true' observation compared to the second situation. This property of MAs is not shared by robust estimation methods such as M-estimation.

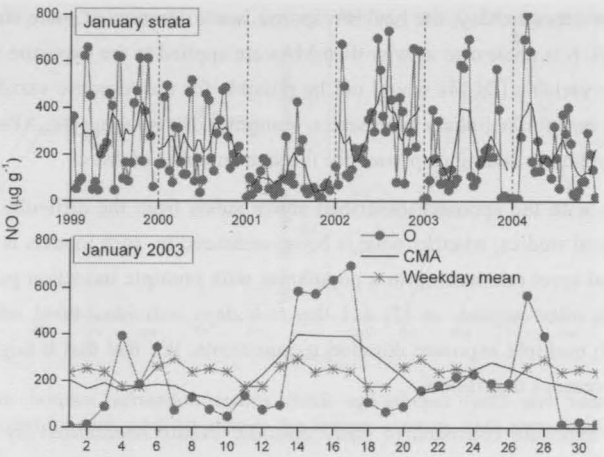



Figure 2.1 Sequential plot of daily levels of NO in the January stratum and close-up for the year 2003. Both the original data (O) and the CMA data are plotted as well as the mean concentration for each weekday type.

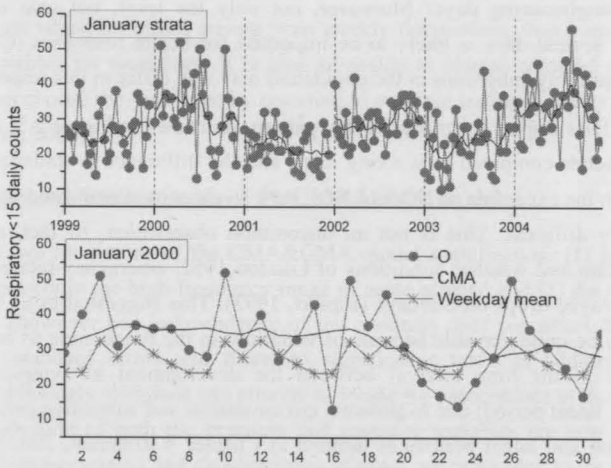



Figure 2.2 Sequential plot of daily levels of selected hospital admission counts of Respiratory>15 in the January stratum and close-up for the year 2000. Both the original data (O) and the CMA data are plotted as well as the mean count for each weekday type.

### 2.4.1.2 Response duration or exposure duration?

A recent trend in acute epidemiological studies of air pollution has been to model single-day health responses with exposures of multiple days, either as MAs or DLMs. Such asymmetrical models are based on the assumption that the associations occur on a 1 day-to-1 day basis at the individual level, but owing to the likely existence of multiple susceptibility subgroups in large populations, different time-intervals (lags) are allowed to exist between the single-day's exposure and the single-day's response (Roberts, 2005). This reasoning implies

that for a given exposure day, the health response has a duration of more than 1 day at the population level. It is unclear to us why then MAs are applied to the exposure variable instead of the response variable (DLMs would not be possible for the response variable). As Lipfert (1993) pointed out, for stationary time-series, summing or averaging the APs over a number of days is equivalent to adding responses for the same number of days.

Our scepticism with the approach described above stems from the difficulty in knowing, at least in ecological studies, whether what is being measured by such models is: (1) a 1 day-to-1 day individual-level relationship in a population with multiple induction periods (response duration), as is often argued; or (2) a 1 day-to-k days individual-level relationships in a population with multiple exposure duration requirements. We feel that it might be a mixture of both for the reasons that follow.

Firstly, if AP exposure consisted of daily episodic events intercalated by absent or sub-threshold exposures, the idea that exposures on a single day could elicit health responses on any number of days would be straightforward. However, AP exposure is permanent and usually displays a healthy degree of autocorrelation in the short term. In this scenario, is it realistic to expect that we can distinguish the effect of 1 day's exposure from the effect of exposure on neighbouring days? Moreover, not only the level, but also the duration of exposure over several days is likely to be important for health responses (Cox, 2000), and different susceptibility subgroups in the population may also differ in this respect.

Secondly, our HAs display a curious weekly pattern. As shown in Figure 2.3, HAs counts are lower on weekends compared to working days, and the difference is statistically significant ($p(t)<5\%$).With the exception of NO and $NO_2$, APs levels on weekends and working days are not statistically different. This is not an uncommon observation. In fact, even under the extreme pollution and weather conditions of London 1952, emergency-related HAs, but not mortality, displayed drops on Sundays (Lipfert, 1993). This suggested to us that the date of admission may be quite versatile because of variations in the functioning of hospital services and variations in the time interval between the development of symptoms and actual admission (i.e. latent period) due to personal circumstances and subjective perception of well-being (induction and latent periods as defined in Chapter 4 Rothman, 2002). The weekend effect could be the clearest manifestation of the impact of such factors but there is no reason why it should not occur to some extent every day. This further suggested to us that, even if the induction periods were known, such factors could lead to unpredictable shifts in admission dates of at least 1–3 days, which could effectively blur any attempts to attribute a particular day's admission to a particular day's exposure.

In this context, the advantage of the CMA&CMA model specification is that it bypasses the need to make assumptions regarding the underlying time-scale of associations and what they represent (e.g. induction periods or exposure duration?), which may be an impossible endeavour in ecological studies involving such large heterogeneous populations and such long follow-up periods. It does so by allowing each daily observation to embody the overall weekly context in which both exposures and responses arise.

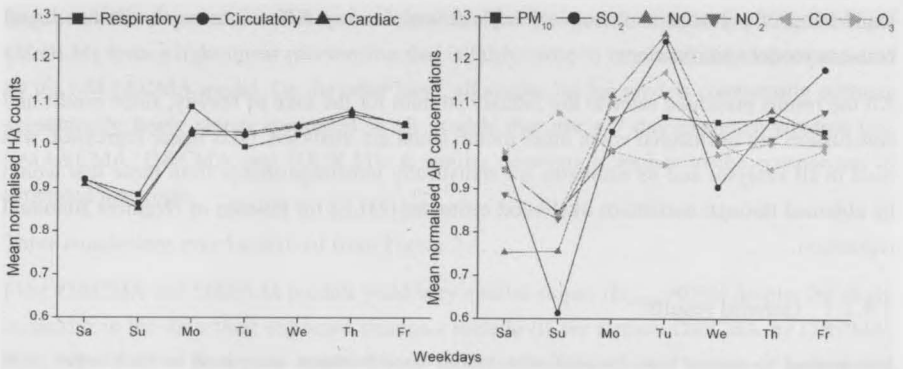



Figure 2.3 Mean hospital admission counts (total age-groups) (left) and mean air pollutant concentrations (right) for each weekday type in the January stratum, both normalised to mean one to help visualisation.

A MA window of 7 days was selected because: (1) it can give considerable flexibility for varying exposure duration requirements and latent/induction periods; (2) 7 days is still considered an acute time-scale for associations (WHO, 2005) and (3) MAs of 7 days in both APs and HAs can adjust for biases arising from weekly fluctuations, thus bypassing the need for dummy variables for weekdays. It is also advisable to choose a model specification a priori rather than choose the 'best' one by resorting to multiple testing (Lumley and Sheppard, 2000; Smith et al, 2000).

#### 2.4.1.3 Anticipated disadvantages of the CMA&CMA model

We anticipated two limitations of the CMA&CMA model specification: (1) loss of relevant information precisely in the high-frequency range of acute effects; and (2) the introduction of autocorrelation (however non-independence of the residuals does not affect the coefficients but only their standard errors and therefore significance test). In addition, this model specification deliberately abandons any attempt to locate the associations at the daily level, as each daily observation of both the exposure and response variables are now averages of 7 days. Having these advantages and disadvantages in mind, we proceeded to the comparison of the CMA&CMA model with the most widely used model specifications in the field.

### 2.4.2 Comparison of model specifications

Studies of the acute health effects of air pollution have used a wide range of specifications and time spans to represent exposure windows and lag-intervals. We compare most such specifications by keeping the span of the exposure window and lag-interval fixed at 7 and 0 days, respectively; except for the single-day lag model specification where the exposure window is of 1 day and the lag interval varies from 0 to 6 days.

This section begins with a comparison of the intercepts and slopes obtained from each model specification with those obtained from the CMA&CMA model. Then we proceed with an

exploration of potential mathematical explanations for any differences or similarities found between model specifications.

All the results presented refer to the January stratum for the sake of brevity, since results and conclusions are unchanged when other month strata are analysed. OLS linear regression was used in all analyses and its estimates are statistically indistinguishable from those that would be obtained through maximum likelihood estimates (MLE) for Poisson or Negative Binomial regression.

### 2.4.2.1 General results

We wished to assess how the different model specifications compared to each other with respect to the intercept and slope. In order to do so, we compared each model specification at a time with the CMA&CMA model (reference model). The comparisons were performed by a simple linear regression where the y-variable was the intercepts or slopes obtained with the CMA&CMA model and the x-variable was the intercepts or slopes of one of the other alternative model specifications. From this regression, we calculated the degree of linear agreement ($R^2$) and slope between the intercepts ($B_{int}$) and slopes ($B_{slopes}$) obtained from each model specification relative to those obtained from the CMA&CMA model. Only those HA-AP relationships that showed a significant (1%) relationship for both model specifications being compared were considered in these regressions. The results of these comparisons are displayed in Figure 2.4.

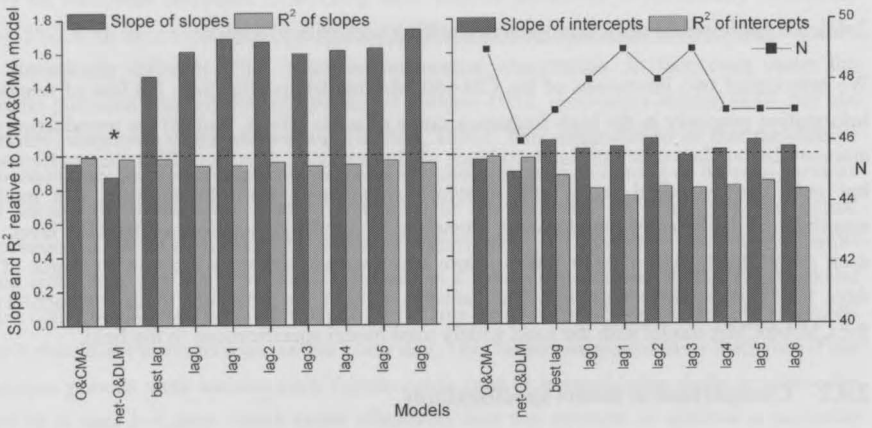



Figure 2.4 Relationship (evaluated by the $R^2$ and slope) between the slopes (left) and the intercepts (right) estimated by the CMA&CMA model and each of the alternative model specifications displayed on the x-axis. For models that included a weekend dummy variable, the mean intercept was used in the comparisons. The values of the slopes on the two graphs should be read as: increase (>1) or decrease (<1) of the slope or intercept of the CMA&CMA model per unit increase in the slope or intercept of a model on the x-axis. Only those HA-AP relationships that were statistically significant at 1% for each pair of models being compared were considered in the graphs; the number (N) of such HA-AP relationships is displayed in the right graph (total number of possible HA-AP relationships is 72).

In general, the slopes obtained from models with a 7 day exposure window (i.e. O&CMA and O&DLM) form a tight linear relationship that is fairly close to one, with the slopes estimated by the CMA&CMA model. On the other hand, all single-day lag models consistently estimate substantially lower slopes compared to all models that use a 7 day exposure window (i.e. CMA&CMA, O&CMA and O&DLM). A similar description applies to the comparison of estimated intercepts.

Three conclusions may be derived from Figure 2.4.

First, O&CMA and O&DLM models yield very similar slopes ($B_{slopes}$=0.92) despite the slight mismatch in the days their exposure windows include (if we replace O&CMA by O&PMA, $B_{slopes}$=0.96). This finding is intuitive and reflects the equivalence between 'aggregating estimates or estimating aggregates' (Cox, 2000). Nevertheless, some authors appear not to have recognised this equivalence or appreciated its implications (e.g. Roberts, 2005; Braga et al, 2001).

Second, an exposure window of multiple days expressed either as a MA or as the net-slope of DLMs consistently yields substantially higher slopes than a single-day exposure window. This finding has been reported in numerous studies and in the next section we will attempt to find a mathematical explanation for it.

Finally, what is new in Figure 2.4 is that the concomitant averaging of the HAs in the CMA&CMA model, in addition to the averaging of the APs, does not lead to substantial changes in estimates compared to O&CMA and O&DLM models ($B_{slopes}$=0.95 and 0.87, respectively). These results are qualitatively unchanged, when we replace CMA&CMA and O&CMA models by FMA&PMA and O&PMA models, respectively (not shown).

#### 2.4.2.2 Smoothing the exposure window: comparison of single-day lags with O&CMA models

As described above, on average, slopes estimated from models that use a 7 day exposure window (CMA&CMA, O&CMA and O&DLM) are substantially higher than those obtained from single-day exposure windows (Figure 2.4). Because CMA&CMA models yield slopes that are very close to those obtained from O&CMA and O&DLM models, it must be concluded that the increase in slopes is mostly due to the averaging of the APs.

The tendency for models to display larger slopes as the span of the exposure window increases is a common observation in the literature (Goodman et al, 2004 and references therein). The prevailing explanation for this phenomenon is conceptual: longer exposure windows are able to capture single-day responses due to single-day exposures at multiple lag-intervals, where the latter are thought to reflect multiple induction periods in a heterogeneous population (Roberts, 2005; Goodman et al, 2004). This explanation has been the major justification for causal studies to report the optimum exposure window, i.e. the one that gives the largest slope.

We feel there are two major problems with this explanation. First, it is questionable whether it can be proven solely on the basis of ecological data. Secondly, we also find it questionable whether it is sound to compare models involving different exposure windows because: (1) although they pertain to the same exposure, averaging (CMA or DLM) leads to different datasets with different statistical properties; and (2) the associations may reflect the same phenomenon occurring at different time-scales, which may be equally valid but not directly comparable.

Assuming it is indeed sound to compare models with different exposure specifications we hypothesised whether mathematical explanations rather than the prevailing explanation described above might contribute to the fairly systematic finding that longer exposure windows yield larger slopes than shorter ones. Because O&CMA and O&DLM models yield similar estimates, we focused the comparison on O&CMA and single-day lag models. From a mathematical point of view, a MA is a smoothed version of the original dataset from which it was calculated; therefore, the statistical properties of the MA dataset and the original dataset are rather different. Therefore, we investigated whether changes in the distribution (skewness and kurtosis) as well as in influential values and outliers between the two datasets might be associated with the different slopes. Our observations and a simulation were not able to pinpoint one of these distributional properties as a cause for the difference in slopes (not shown).

In a similar way, we hypothesised whether the reduction in the variance of the APs, induced by the O&CMA model, could have played a role. If we consider the simple case of just one independent variable, the slope is calculated by $b = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sum(x_i - \bar{X})^2} = \frac{\sum(x_i y_i) - \bar{X}\bar{Y}}{\sum(x_i x_i) - \bar{X}\bar{X}}$. This formula suggests that a reduction in the variance of the independent variable may reduce the denominator to a greater extent than the numerator. However, it is difficult to make general statements because the extent to which the decrease in covariance might be due to the simple shrinkage of the x-values or due to an effective change in the structure of the variation of x relative to y is hard to pin down. Moreover, any generalisations are complicated by the fact that both the denominator and numerator are subtractions and a small difference between two large numbers can be very unstable.

Finally, we performed a comparison between the O&CMA and the single-day lag model (lag0) in a different way to the comparison performed in Figure 2.4.We calculated the difference in mean intercept between the two models and the difference in slope between the same two models and then scatter-plotted the differences. Figure 2.5 shows the resulting relationship. On the right hand-side of the y-axis, HA-AP relationships have a negative slope for both models being compared, whereas on the left hand-side they have positive slopes. The graph indicates that for most HA-AP relationships the greater the difference in mean intercept between the two models, the greater the difference in slope between the two models, and vice-versa. Why should such a relationship exist? A change in slopes does not necessarily have to be accompanied by a change in intercept and vice-versa.

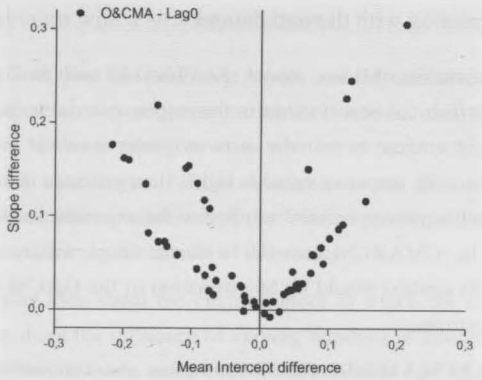



Figure 2.5 Relationship between the difference in mean intercept between the O&CMA model and the lag0 model (i.e. x-axis=mean intercept of O&CMA - mean intercept of lag0) with the difference in slope between the same two models (i.e. y-axis=slope of O&CMA - slope of lag0), for each HA-AP relationship that was significant for both models. The HA-AP relationships to the right side of the y-axis have negative slopes for both models whereas those to the left side have positive slopes.

In summary, this section attempted to explore potential mathematical explanations for the fact that models containing a 7 day exposure window display larger slopes than models with a 1 day exposure window. We have been unsuccessful at finding a definite explanation; but this may be due to the fact that we looked at the role of a single factor at a time (e.g. variance reduction or reduction of outliers or skewness) when it is likely that several factors act simultaneously to produce the change in slope. This issue should warrant further investigation in the future.

#### 2.4.2.3 Smoothing the response window: comparison of CMA&CMA and O&CMA

The slopes estimated by the CMA&CMA models are slightly lower but very close to those obtained from O&CMA models ($B_{slopes}$=0.95) (Figure 2.4). If we compare these models with single-day lag models, it is clear that averaging the HAs leads to a much smaller change in the slopes compared to smoothing just the APs. This result suggests two preliminary conclusions. Firstly, the original HAs dataset does not appear to contain influential values relative to the CMA HA dataset, because if it did they would be smoothed by the CMA which in turn would lead to more substantial changes in slopes. Secondly, the original HAs dataset appears to contain superfluous variation with little informative value (noise), since the variability lost by averaging the HAs into CMA does not impact the slopes to any great extent while substantially increasing the precision.

### 2.4.3 Evaluation of the CMA&CMA model

In this section we perform a more detailed comparison of the CMA&CMA model with the O&CMA model in terms of their sensitivity to noise and extreme observations in the response variable and to the deletion of single observations in the dataset.

### 2.4.3.1 Robust regression with the real dataset

Many epidemiological studies that use model specifications such as O&CMA have used M-estimation to handle influential observations in the response variable (e.g. Samoli et al, 2001; Schwartz, 2000b). It is unclear to us why authors prefer a robust regression method that targets extreme values in the response variable rather than methods that target extreme values in both the response and exposure variables or just in the exposure variable. Nevertheless, our aim is to compare the CMA&CMA model with the most widely used methods in the literature, which in this context would be M-estimation of the O&CMA rather than the OLS used in the previous sections.

Since both the CMA&CMA model and M-estimation share an ability to handle extreme values in the response variable, it was hypothesised that the slopes obtained from CMA&CMA models and O&CMA models might become more similar if the latter is estimated by M-estimation. Despite the apparent absence of extreme values in our HA dataset, it does appear to have a noisy pattern (Figure 2.2) which M-estimation may adjust for.

It is rarely stated in the literature which M-estimator was used, therefore we opted for using Huber's and Tukey's Biweight estimators in R. Figure 2.6 shows the relationship between the slopes obtained from the CMA&CMA model (y-axis) and the slopes obtained from the O&CMA model estimated by each robust estimator (x-axis), across statistically significant (1%) HA-AP relationships. It can be concluded that the slopes obtained by the CMA&CMA model estimated by OLS is fairly similar to the slopes obtained by the O&CMA models estimated by either robust estimator ($B_{slopes}$=0.95 for either robust estimator). The fact that the slopes obtained by O&CMA estimated by OLS are very similar to those of O&CMA estimated by the robust estimators ($B_{slopes}$=0.99 for Huber's and 0.97 for Tukey's) suggests that the HAs data is fairly devoid of influential values.

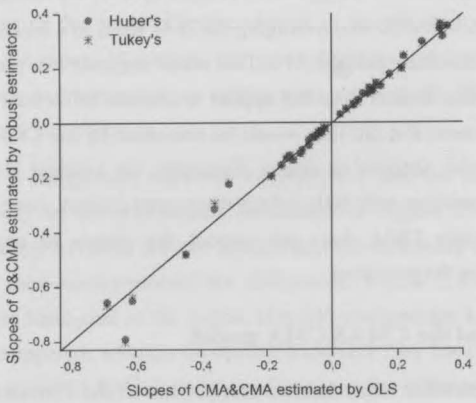



Figure 2.6 Relationship between the slopes obtained from the CMA&CMA model estimated by OLS (y-axis) and the slopes obtained from the O&CMA model estimated by either Huber's or Tukey's estimator (x-axis). Only those HA-AP relationships that were statistically significant at 1% for the CMA&CMA model were considered in the graph. The value of the slope and $R^2$ of both relationships are displayed on the graph.

### 2.4.3.2 Robust regression with a simulated dataset

As explained before, both the CMA&CMA model and M-estimation are able to handle extreme values and noise in the response variable. However, this ability differs in one important aspect: while MAs are sensitive as to whether extreme values are consecutive or isolated in time (Figure 2.1 and Figure 2.2), M-estimation is not. This ability is of importance when evaluating whether extreme values might be 'real' or errors and to what extent we are willing to allow them to influence the results of the analysis.

In order to have a clearer idea about the circumstances in which the results of the previous section might arise and about the influence of varying numbers of consecutive extreme values in the slopes obtained from the two model specifications (CMA&CMA and O&CMA) and three estimation methods (OLS, Huber's and Tukey's), we performed a simple simulation. A dataset consisting of 200 observations was used to create eight scenarios where: zero to seven consecutive observations were made extreme (the extreme values were approximately 2.5 times higher than the mean value of the 200 observations without extreme values). Figure 2.7 displays the slopes estimated by the two model specifications and three estimation methods, across the eight scenarios. The two robust estimators reveal a striking difference in their sensitivity to the number of extreme values: O&CMA_Tukey's estimator being the least sensitive whereas O&CMA_Huber's estimator occupies an intermediate position between O&CMA_Tukey's and O&CMA_OLS.

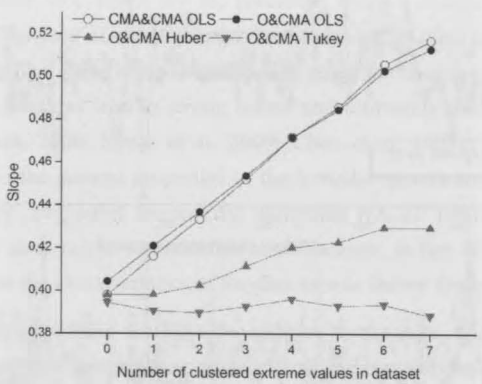



Figure 2.7 Slopes estimated from a simulated dataset consisting of 200 observations where 0 to 7 consecutive observations were made extreme (x-axis). The extreme observations were approximately 2.5 times higher than the mean of the dataset with 0 extreme values. The slopes for these eight scenarios are presented for the CMA&CMA model estimated by OLS, and for the O&CMA model estimated by OLS, Huber's and Tukey's estimator.

It is interesting to note that in the scenario where the response variable has no extreme values, the slope estimated by CMA&CMA is very similar to that obtained from O&CMA with M-estimation, or OLS estimation. This is in fact what was found with our real HA-AP dataset in the previous section. However, when the number of consecutive extreme values is low (say

less than 5), the slope obtained from CMA&CMA gradually approaches the slope obtained from O&CMA_OLS; whereas, the robust estimators give comparatively lower and lower slopes. Finally, as the number of consecutive extreme values increases the slope of the CMA&CMA model gradually becomes larger and larger than the slope of O&CMA_OLS; whereas, the robust estimators remain conservative. The fact that CMA&CMA models may give slopes that are larger than those obtained from O&CMA_OLS indicates that when extreme values are abundant instead of smoothing, the CMA&CMA 'broadens' extreme values over neighbouring observations.

### 2.4.3.3 Robustness of estimates

In a final comparative analysis we tested an important property of any model specification: that its estimates are robust to small changes in the dataset. One way of investigating this is to remove a single observation across the range of available observations and see how that impacts the slope. This is exemplified for two HA-AP relationships in Figure 2.8 where it can be observed that the slopes obtained by the CMA&CMA model estimated by OLS shows much smaller fluctuations than the slopes obtained by the O&CMA model, whether the latter is estimated by OLS or by Tukey's estimator.

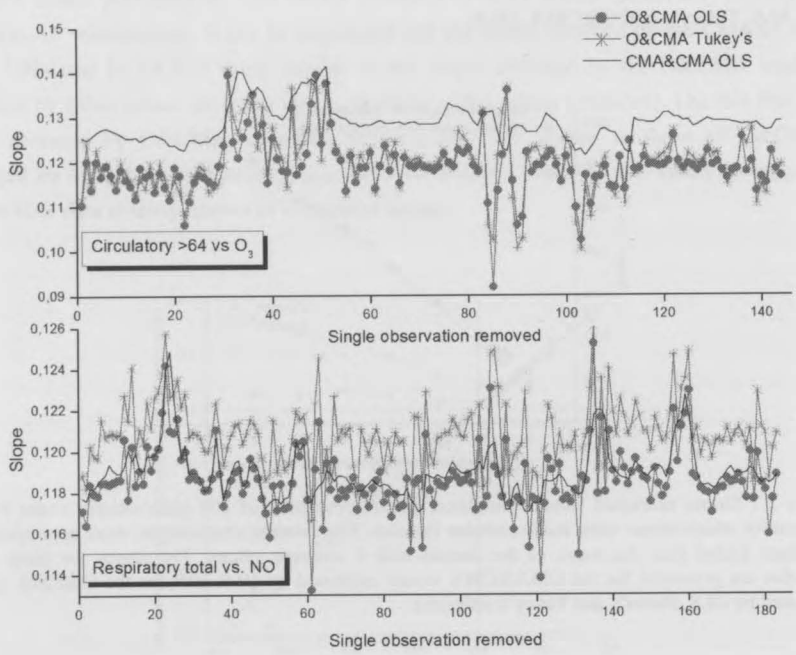



Figure 2.8 Change in slope when a single observation is removed across the range of observations available in the dataset (x-axis). This is exemplified for two HA-AP relationships: Circulatory>64 with O3 (top) and respiratory total with NO (bottom). The slopes were obtained from the CMA&CMA model estimated by OLS and from the O&CMA model estimated with Tukey's estimator.

## 2.5 Discussion

The new model specification, CMA&CMA, was conceived a priori on the basis of considerations regarding data quality and characteristics, the ecological design of the study and the processes that might interfere with acute health effects. This model was then compared with the most widely used model specifications (namely: single-day lags, O&CMA and O&DLM models) and estimation methods (M-estimation) in the literature, by taking as a sample the 72 HA-AP relationships available. Results were presented only for the January stratum since results were unchanged for the other month strata. All analyses were performed by linear regression as this method leads to the same estimates (after an approximation) as Poisson regression and our aim is a comparison of methods, not causal inference.

To the best of our knowledge only one article has reported the use of MAs in both the response and exposure variable (Schwartz, 2000a) but the reasons for doing so and its consequences relative to other model specifications were not made explicit. Apart from Roberts (2005) who proposed the use of moving total counts of mortality as a substitute for MAs in cases where PM measurements are not available daily, no other authors have mentioned the use of aggregated response variables.

Three warning words must be given before proceeding. First, we have compared statistics from different model specifications without knowing which (if any) is the most accurate; the only criteria we are able to evaluate are the precision, relative change in estimates and the latter's robustness in the face of small changes to the dataset. It must be emphasised that, in the absence of independent experimental evidence to guide the choice of model specification, statistical-based decisions can lead to strong biases and ultimately meaningless associations (Lumley and Sheppard, 2000; Smith et al, 2000; Chen et al, 1999). Secondly, the results presented here refer to the general properties of the 5 model specifications across 72 HA-AP relationships, and they may differ for specific individual HA-AP relationships. Thirdly, the results presented here need not be reproducible in all datasets; in fact different results may be expected depending on the characteristics of the data as was shown through a simulation.

Our results show that smoothing the HAs, in addition to the APs, does not lead to loss of information at the fine time-scales where acute effects are conventionally expected to occur. This is evidenced by the fact that the CMA&CMA model yields slopes that are very close to those estimated by the O&CMA model estimated by either OLS, or M-estimation. It can be concluded that, our particular HAs dataset was fairly devoid of influential values, and for this reason both models (O&CMA and CMA&CMA) and both estimation procedures (OLS and M) gave very similar estimates.

However, a simulation revealed that such a result may only occur under certain circumstances. When the response variable has several extreme values that are consecutive in time, CMA&CMA slopes are closer to those of O&CMA_OLS and higher than those of O&CMA_robust. When the response variable has several extreme values that are not

consecutive in time, CMA&CMA slopes are closer to those of O&CMA_robust and lower than those of O&CMA_OLS. This difference emphasises the fact that CMA&CMA model is able to deal with extreme values that are more likely to be ''real'' (due to their being consecutive in time) differently from those that are less likely to be ''errors'' (because they are sporadic), a property that is not shared by M-estimation which down weighs extreme values regardless of their proximity in time.

In what concerns the precision and robustness of the slopes, the CMA&CMA model outperforms the O&CMA model even if the latter is estimated by M-estimation. The increased precision and robustness of the slopes obtained from CMA&CMA models may stem from two sources: the fact that each data point is based on seven observations rather than just one and the introduction of serial correlation as the Durbin-Watson statistic is always inferior to one for the CMA&CMA dataset whereas it is usually very between one and two for the O&CMA dataset. The extent to which each of these factors contributes to the increased precision and robustness of the CMA&CMA models is difficult to disentangle.

It was shown by means of a simulated dataset that different M-estimators differ greatly in their sensitivity to the number of clustered extreme values. While this is not new, the fact that many authors do not report the M-estimator they used (e.g. Samoli et al, 2001; Schwartz, 2000b) allied to the fact that the major statistical packages for robust regression (R and S-PLUS) differ in their default weighing methods, appears to overlook the great impact that this decision may have on effect estimates intended for causal inference. In addition, it is important to consider that, while M-estimators have different sensitivities to the number of extreme values in the response variable they are, contrarily to MAs, completely blind as to whether they are also clustered in time. Both of these properties are important for evaluating, in the absence of additional information, whether extreme values are 'real' or errors.

In conclusion, the CMA&CMA model estimated by OLS has the following advantages for datasets similar to ours: (1) it yields estimates that are substantially more robust and more precise compared to other model specifications and two major M-estimators; (2) MAs are sensitive to the number of clustered extreme values, a property that not all M-estimators share; (3) MAs are sensitive as to whether extreme values are clustered in time, a property that none of the M-estimators share; (4) the CMA&CMA model is easy to use and does not require specialised software; (5) when the window of the CMA&CMA model is of 7 days, it can adjust for systematic weekly oscillations, thus avoiding the need to include additional terms for weekdays, and can adjust for effect modification (which dummy variables cannot). In the case of datasets with different properties from ours, the CMA&CMA model may serve as a tool to assess data properties and quality more closely. We had anticipated two potential disadvantages of the CMA&CMA model. First, the loss of relevant high frequency information due to the smoothing of HAs, but this was shown to be unfounded. The second was the introduction of autocorrelation in the residuals, which is confirmed as the Durbin-Watson statistic is always less than one.

The pursuit of the daily level of analysis in time-series studies appears to have grown out of convention, the availability of data on a daily basis, and the statistical advantages offered by a

large dataset. In this article we have tried to introduce another criteria for the time-unit of analysis, that of data quality. There could be instances when single-day health data at the population level is simply not reliable enough and/or contains no additional information compared to aggregations over more than 1 day, even though we may have strong reasons to believe that associations do occur on a daily basis at the individual level. As shown in this study, the averaging of the HAs in time may in fact be advantageous as it increases the precision of the estimates without distortion or loss of the underlying daily signal. However, with such aggregations we lose the ability to locate or attribute the associations with single days and we introduce autocorrelation in the residuals.

# 3 How many samples, where to sample and the quantification of the between-area to within-area variance ratio – a simulation study

*Partly based on the review article:*
*"Is there a future for biomonitoring of elemental air pollution?*
*A review focused on a larger-scale health-related (epidemiological) context".*
*Wolterbeek HTh, Sarmento SM & Verburg TG.*
*Journal of Radioanalytical and Nuclear Chemistry (2010) 286, 195-210.*

## 3.1 Abstract

The present study presents practical recommendations for obtaining accurate and precise estimates from sampling surveys, especially, but not limited to, spatial ones. The recommendations target mostly the question of how many samples and where to sample. Also investigated are the consequences of less than optimal sample sizes for the accuracy and precision of sampling survey's estimates and its implications for the survey's signal-to-noise ratio.

The minimum sample size (MSS) required to estimate the average and the variance of populations with characteristics such as, but not restricted to, those found in geographical surveys of environmental parameters are provided. A simulation procedure known as sampling-without-replacement is shown to be a viable alternative to sample-size formulas which, contrarily to the latter, can calculate the MSS required to estimate the variance of skewed populations and virtually any statistic from virtually any type of distribution. The simulation procedure also enabled the identification of the population's characteristics that are determinant for the MSS required to estimate the average and the variance.

Three of the most commonly used probability-based sampling methods were tested with regards to their ability to estimate several parameters of a population with spatial structure. For a fixed sample size, the estimates obtained with systematic-grid sampling present the lowest sampling variability for a wider range of statistics, including tail-values; however because this type of sampling is often impractical, random-within-blocks sampling appears to be a feasible second best choice.

The sampling density of the survey and of the sampling sites composing the surveys were investigated in terms of how they impact the accuracy and precision of both the survey's and the sampling sites' estimates. The higher the moment of the statistic being estimated the more it is vulnerable to being smoothed out by aggregation and to sampling variability. The local variance (i.e. variance within the survey's sampling sites) gives a measure of the uncertainty of the survey's estimates and thus need to be estimated accurately in order to be kept low. It was found that less than the optimal number of samples tends to underestimate local variances

and thus inflate the signal-to-noise ratio. The intimate relationship between local variances and survey's sampling density is discussed as is their relevant to the signal-to-noise ratio.

The realisation that local variances need to be estimated as rigorously as the survey's variance in order to obtain the latter's uncertainty and to calculate a sound signal-to-noise ratio greatly increases the number of samples that need to be sampled and analysed individually in a single survey. Alternative or complementary methods to estimate or correct local variances with a lower number of samples/analyses should be the target of further research in the future.

## 3.2 Introduction

In many realms of science, and environmental science in particular, it is not unusual for sampling surveys to measure multiple parameters simultaneously. In the specific context of atmospheric biomonitoring (Wolterbeek et al, 2010), multi-analytical nuclear techniques (Freitas et al, 2000) and monitoring networks of regulated air pollutants, this is routine. While the convenience and cost reduction of this approach are certainly undeniable, the number of samples in such surveys is unlikely to be equally adequate for all parameters being estimated. To have a rough idea, Tables 3.1-3.3 shows descriptive statistics of chemical elements measured in country-wide lichen and moss surveys (n~110-290) in Portugal (Freitas et al, 1999, reproduced with permission), the Netherlands (Kuik & Wolterbeek, 1995, reproduced with permission) and Slovenia (Jeran et al, 1996, 2003, reproduced with permission). The diversity in estimates across the chemical elements is striking, with relative standard deviations (RSD) ranging from 20% to 200%, skewness escalating up to 10 and kurtosis up to 150. This diversity in the value of sample estimates suggests that the respective population's parameters are also very diverse across chemical elements and therefore different chemical elements would require different sample sizes in order to be equally accurately and precisely estimated.

Table 3.1 Descriptive statistics for chemical elements concentrations ($\mu g\ g^{-1}$) measured in the lichen (*Parmelia sulcata*) sampling survey carried out in Portugal in 1993 (Freitas et al, 1999; reproduced with permission). RSD-relative standard deviation; LC-local variance.

| | N | Average | SD | RSD(%) | LV(%) | Med | Skew | Kurt |
|---|---|---|---|---|---|---|---|---|
| Al | 294 | 5417 | 3478 | 64.2 | 28.3 | 4645 | 2.82 | 12.75 |
| As | 297 | 2.15 | 2.90 | 135.0 | 19.9 | 1.35 | 5.77 | 43.01 |
| Ba | 246 | 33.28 | 20.05 | 60.3 | 16.1 | 29.60 | 2.10 | 7.43 |
| Br | 296 | 22.62 | 11.35 | 50.2 | 20.3 | 19.85 | 1.72 | 5.52 |
| Ca | 251 | 6497 | 3474 | 53.5 | 24.7 | 5820 | 2.12 | 6.96 |
| Cl | 286 | 1407 | 983 | 69.9 | 23.8 | 1245 | 10.11 | 140.18 |
| Co | 292 | 0.779 | 0.477 | 61.2 | 23.4 | 0.670 | 1.90 | 4.39 |
| Cr | 297 | 5.47 | 3.62 | 66.3 | 24.9 | 4.69 | 3.84 | 25.10 |
| Cs | 295 | 0.642 | 0.507 | 78.9 | 24.0 | 0.514 | 3.15 | 13.82 |
| Eu | 285 | 0.185 | 0.103 | 56.1 | 27.7 | 0.163 | 1.82 | 4.70 |
| Fe | 293 | 2110 | 1181 | 56.0 | 23.2 | 1920 | 1.64 | 3.55 |
| K | 296 | 5253 | 1554 | 29.6 | 12.3 | 5010 | 0.91 | 1.18 |
| La | 297 | 3.02 | 1.95 | 64.6 | 22.0 | 2.56 | 2.14 | 6.12 |
| Mg | 284 | 1903 | 764 | 40.1 | 30.8 | 1755 | 1.04 | 1.52 |
| Mn | 295 | 52.06 | 36.28 | 69.7 | 17.3 | 46.40 | 10.39 | 146.22 |
| Na | 297 | 589 | 355 | 60.2 | 20.3 | 492 | 2.05 | 6.77 |
| Rb | 296 | 15.85 | 9.08 | 57.3 | 22.1 | 13.60 | 2.13 | 8.83 |
| Sb | 287 | 0.343 | 0.490 | 142.6 | 21.9 | 0.220 | 8.89 | 104.86 |
| Sc | 293 | 0.656 | 0.407 | 62.0 | 23.3 | 0.562 | 1.99 | 5.31 |
| Se | 295 | 0.403 | 0.153 | 38.0 | 20.8 | 0.373 | 1.49 | 3.73 |
| Sm | 297 | 0.448 | 0.256 | 57.2 | 30.9 | 0.392 | 1.77 | 5.10 |
| V | 286 | 17.16 | 27.04 | 157.5 | 20.3 | 11.45 | 9.69 | 121.01 |

Table 3.2 Descriptive statistics for chemical elements concentrations (μg $g^{-1}$) measured in the moss (*Pleurozium schreberi*) sampling survey carried out in the Netherlands in 1992 (Kuik & Wolterbeek, 1995; reproduced with permission). RSD-relative standard deviation; LC-local variance.

| | N | Average | SD | RSD(%) | LV(%) | Med | Skew | Kurt |
|---|---|---|---|---|---|---|---|---|
| Al | 109 | 870 | 514 | 59.13 | 18.27 | 731.8 | 2.88 | 10.27 |
| As | 109 | 0.435 | 0.157 | 36.04 | 22.40 | 0.3987 | 1.33 | 2.88 |
| Ba | 109 | 27.1 | 10.3 | 37.86 | 36.91 | 25.89 | 0.52 | -0.33 |
| Br | 109 | 7.05 | 3.37 | 47.87 | 19.82 | 6.089 | 2.82 | 12.69 |
| Ca | 109 | 2489 | 738 | 29.66 | 13.26 | 2416 | 3.28 | 20.68 |
| Cd | 109 | 3.83 | 0.82 | 21.32 | 16.05 | 3.766 | -0.12 | 1.67 |
| Ce | 109 | 1.38 | 1.38 | 100.34 | 29.00 | 0.9911 | 4.66 | 26.37 |
| Cl | 109 | 433 | 200 | 46.22 | 21.10 | 407.1 | 1.11 | 1.95 |
| Co | 109 | 0.372 | 0.182 | 48.97 | 19.29 | 0.3068 | 1.34 | 1.98 |
| Cr | 109 | 5.73 | 4.21 | 73.44 | 49.51 | 4.223 | 1.80 | 4.23 |
| Cs | 109 | 0.225 | 0.095 | 42.25 | 26.35 | 0.2008 | 1.61 | 4.39 |
| Cu | 109 | 22.3 | 5.7 | 25.65 | 18.22 | 23.99 | -0.51 | -0.61 |
| Eu | 109 | 0.024 | 0.015 | 63.32 | 22.43 | 0.01909 | 2.83 | 8.73 |
| Fe | 109 | 739 | 356 | 48.11 | 15.88 | 645.2 | 3.42 | 17.53 |
| Hf | 109 | 0.327 | 0.782 | 238.79 | 51.75 | 0.1159 | 4.49 | 20.37 |
| Hg | 109 | 0.172 | 0.036 | 21.19 | 17.59 | 0.1675 | 0.68 | 1.17 |
| I | 109 | 3.65 | 1.28 | 35.18 | 24.42 | 3.465 | 1.87 | 5.86 |
| K | 109 | 5337 | 1217 | 22.81 | 14.53 | 5072 | 0.76 | -0.15 |
| La | 109 | 0.742 | 0.785 | 105.77 | 23.31 | 0.5521 | 5.36 | 34.77 |
| Lu | 109 | 0.015 | 0.009 | 57.98 | 24.42 | 0.01369 | 3.46 | 14.16 |
| Mg | 109 | 1344 | 296 | 22.02 | 18.55 | 1308 | 0.11 | -0.38 |
| Mn | 109 | 222 | 178 | 80.12 | 21.24 | 162 | 2.52 | 7.43 |
| Mo | 109 | 2.76 | 3.33 | 120.64 | 31.76 | 2.167 | 6.03 | 36.72 |
| Na | 109 | 421 | 230 | 54.77 | 13.34 | 378.7 | 1.80 | 3.78 |
| Nd | 109 | 3.14 | 0.74 | 23.55 | 21.92 | 3.038 | 0.06 | 0.95 |
| Ni | 109 | 15.4 | 6.0 | 38.86 | 27.96 | 15.03 | 3.99 | 30.74 |
| Rb | 109 | 18.8 | 7.5 | 40.03 | 19.18 | 18.96 | 0.37 | 0.14 |
| Sb | 109 | 0.601 | 0.241 | 40.03 | 18.39 | 0.5509 | 1.23 | 2.42 |
| Sc | 109 | 0.154 | 0.122 | 79.54 | 19.23 | 0.122 | 3.71 | 14.77 |
| Se | 109 | 0.498 | 0.140 | 28.16 | 23.88 | 0.5014 | 0.85 | 2.26 |
| Sm | 109 | 0.088 | 0.092 | 104.02 | 30.19 | 0.06258 | 3.56 | 14.01 |
| Sn | 109 | 7.71 | 1.44 | 18.71 | 13.17 | 7.627 | 0.00 | 3.06 |
| Sr | 109 | 28.4 | 5.7 | 20.18 | 13.98 | 27.07 | 1.38 | 2.07 |
| Tb | 109 | 0.037 | 0.011 | 31.34 | 30.67 | 0.03826 | -0.11 | -0.13 |
| Th | 109 | 0.163 | 0.170 | 104.48 | 27.80 | 0.116 | 4.62 | 26.79 |
| Ti | 109 | 203 | 274 | 134.95 | 36.73 | 117.4 | 3.41 | 11.65 |
| U | 109 | 0.481 | 0.097 | 20.20 | 15.81 | 0.4782 | 0.37 | 0.49 |
| V | 109 | 4.68 | 1.15 | 24.59 | 14.46 | 4.538 | 1.18 | 3.11 |
| W | 109 | 0.545 | 0.224 | 41.19 | 30.25 | 0.557 | 0.02 | -0.79 |
| Yb | 109 | 0.073 | 0.056 | 77.24 | 30.17 | 0.05441 | 3.80 | 16.29 |
| Zn | 109 | 65.0 | 25.0 | 38.42 | 13.46 | 68.51 | 0.49 | 1.08 |

Table 3.3 Descriptive statistics for chemical elements concentrations (μg $g^{-1}$) measured in the lichen (*Hypogymnia physodes*) sampling survey carried out in Slovenia in 2001 (Jeran et al, 1996, 2003; reproduced with permission). RSD-relative standard deviation; LC-local variance.

| | N | Average | SD | RSD(%) | LV(%) | Med | Skew | Kurt |
|---|---|---|---|---|---|---|---|---|
| As | 190 | 0.5230 | 0.1623 | 31.03 | 14.69 | 0.4941 | 1.51 | 4.69 |
| Au | 188 | 0.0038 | 0.0084 | 224.12 | 62.48 | 0.0013 | 4.46 | 22.07 |
| Ba | 190 | 29.75 | 48.50 | 163.01 | 29.86 | 22.00 | 9.89 | 116.21 |
| Br | 190 | 11.40 | 3.91 | 34.27 | 16.35 | 10.55 | 1.21 | 2.49 |
| Ca | 190 | 23027 | 11893 | 51.65 | 34.36 | 21254 | 0.78 | 0.89 |
| Cd | 179 | 0.7616 | 0.4603 | 60.44 | 25.36 | 0.6174 | 2.06 | 4.63 |
| Ce | 190 | 1.638 | 0.707 | 43.16 | 19.58 | 1.510 | 2.37 | 11.27 |
| Co | 187 | 0.3803 | 0.1961 | 51.58 | 21.78 | 0.3379 | 2.61 | 9.36 |
| Cr | 190 | 3.712 | 4.342 | 116.98 | 16.76 | 2.834 | 6.55 | 46.43 |
| Cs | 190 | 0.2522 | 0.2206 | 87.44 | 26.37 | 0.1984 | 4.29 | 22.11 |
| Fe | 190 | 766.1 | 309.8 | 40.44 | 18.45 | 684.4 | 2.05 | 7.50 |
| Hf | 190 | 0.1081 | 0.0577 | 53.41 | 23.66 | 0.0944 | 2.06 | 6.24 |
| Hg | 187 | 0.0578 | 0.0221 | 38.19 | 22.82 | 0.0548 | 0.72 | 0.78 |
| K | 190 | 3804 | 992 | 26.08 | 12.61 | 3738 | 0.49 | -0.12 |
| La | 190 | 0.7760 | 0.3341 | 43.05 | 19.78 | 0.7147 | 2.56 | 13.43 |
| Mo | 190 | 0.2657 | 0.2240 | 84.32 | 20.47 | 0.2266 | 5.92 | 40.27 |
| Na | 190 | 128.7 | 50.6 | 39.30 | 17.79 | 115.5 | 1.64 | 3.64 |
| Nd | 190 | 0.8164 | 0.5881 | 72.04 | 25.49 | 0.7350 | 8.01 | 86.57 |
| Rb | 190 | 15.56 | 9.92 | 63.77 | 19.70 | 13.48 | 2.05 | 6.24 |
| Sb | 190 | 0.2306 | 0.1258 | 54.55 | 13.11 | 0.2092 | 6.35 | 58.62 |
| Sc | 190 | 0.2305 | 0.1020 | 44.24 | 20.41 | 0.2084 | 2.44 | 12.01 |
| Se | 190 | 0.2463 | 0.3154 | 128.08 | 29.45 | 0.1877 | 6.59 | 46.20 |
| Sm | 190 | 0.1310 | 0.0543 | 41.47 | 19.63 | 0.1210 | 2.02 | 8.41 |
| Sr | 184 | 27.29 | 13.42 | 49.20 | 25.66 | 25.76 | 1.97 | 9.08 |
| Tb | 185 | 0.0188 | 0.0075 | 39.74 | 20.68 | 0.0182 | 1.57 | 5.01 |
| Th | 190 | 0.1918 | 0.0933 | 48.63 | 20.67 | 0.1729 | 3.05 | 17.18 |
| U | 190 | 0.0688 | 0.0286 | 41.58 | 21.71 | 0.0644 | 1.68 | 5.02 |
| Yb | 190 | 0.0590 | 0.0247 | 41.94 | 20.96 | 0.0554 | 1.58 | 5.14 |
| Zn | 190 | 97.80 | 33.70 | 34.46 | 19.73 | 87.75 | 1.11 | 1.68 |

When planning a sampling survey for a particular individual parameter, it is necessary to adequate the sample size to numerous criteria (e.g.: Chaudhuri & Stenger, 2005; Cochran, 1977), of which we shall address but a few.

The first issue to consider is the purpose of the sampling survey, especially if it is meant for descriptive (e.g.: mapping) or analytical purposes. When the purpose is analytical (e.g.: comparison of pollutant levels across different regions/time, emission source identification, correlation with other variables such as health outcomes), a greater stress on statistical precision and power is required from estimation and the statistics to be estimated may be more elaborate (e.g.: tail-values). This issue is not directly addressed here.

The second issue is the known or suspected population's characteristics (e.g.: distribution, variance, skewness, and kurtosis). A rough idea of the population's characteristics may be derived from previous surveys (e.g.: Tables 3.1-3.3) or small samples (Cochran, 1977), with the eventual assistance of bootstrapping (Wolterbeek & Verburg, 2002). Intuitively, one probably needs more samples when the population's RSD is 100% than when it is 20%, and when its distribution is skewed as opposed to normal.

The third issue is the population's parameter(s) of interest for estimation (e.g.: average, variance, tail-values). Intuitively, one probably needs a larger sample size to estimate the variance than the average. For some purposes, like surveillance and correlation analyses it may be very important to accurately and precisely estimate the tails of the population's distribution.

The fourth issue is the margin of error (ME) and statistical significance level (i.e. statistical power) desired for the estimates. The ME defines how close the estimates should be to the true population's parameter, whereas the statistical significance or power defines the probability of estimating the true population's parameter, within the chosen ME, from a single survey. Obviously the degree of precision and statistical significance required from estimates has a dramatic effect on the sample-size required. Decisions on these two parameters should be mostly driven by the purposes of the sampling survey (see the first issue above) and, as shall be discussed, by the characteristics of the population and of the sampling survey (e.g.: local variance).

The fifth issue is the choice of a sampling method, which defines where the samples are drawn and, in particular, the extent to which their location is random or regular. Depending on the structure of the population (homogeneous or not), different sampling methods may require a different number of samples in order to obtain estimates with the same margin of error and statistical power.

The sixth issue concerns the sampling density of the survey or, equivalently, the aggregation of the population. A decision on this issue should consider both the aggregation required by the purposes of the study (e.g.: accurate representation of the surface of a country or accurate representation of parishes or provinces within a country) and the scales at which the parameters being sampled actually vary in a meaningful way. The current setup of monitoring networks of regulated air pollutants already reflects current knowledge on this issue (e.g.: $O_3$

is measured mostly at non-urban stations and only during day-time). To choose a survey's sampling density is to choose a level of aggregation for the population and thus a scale below which the parameter(s) being sampled are assumed to be homogeneous, which leads us to the next and final issue that shall be addressed here.

The final issue approached in this chapter concerns the survey's signal-to-noise ratio and the estimates required to calculate it (Wolterbeek et al, 1996). Statistical procedures, of which sampling is no exception, are rooted on the assumption that variability can be partitioned. At its most basic level, variability can be partitioned into a deterministic and random component, but the deterministic component can be divided further. With time-series data, for instance, variability can often be clearly divided into time-frequencies (long-term trends, seasons, weekends/weekdays, day/night). Geographical data is not as clear-cut but its variability may be decomposed according to distance (and direction), which interpolation techniques exploit (Cressie, 1993). A signal-to-noise ratio is essentially a ratio between variability of interest and variability of no interest, each defined by the research purposes and/or design of the sampling survey. Wolterbeek et al (1996) proposed that a signal-to-noise ratio for environmental sampling surveys should be expressed as the ratio between the survey's variance and the local variance (SV/LV), i.e. the ratio of the variance between all sampling sites of the survey and the (average) variance between sub-samples within the sampling sites. This is analogous to an ecological epidemiology measure known as the between-area to within-area ratio (B/W) applied mostly to data that is bounded by political regions (e.g.: demographic, socio-economic, health) (Greenland & Robins, 1994a; Salway, 2003; Salway & Wakefield, 2004; Webster, 2007). Both are essentially an F-ratio statistic, and in both cases the aim is to assess the adequateness of the aggregation of the population being sampled. If more variability is present below the aggregation level of the survey (i.e. local or within-area) than above it (i.e. survey or between-area) then the survey's uncertainty is very large, differences between sampling sites/areas are probably erroneous, the survey's quality is poor and the data is useless and misleading for most purposes (Steel et al, 2004; Wolterbeek et al, 2010). Thus one of the fundamental duties of a sampling survey is to monitor the SV/LV ratio and ensure that it is large. This however, means that the number of samples must not only adequately estimate the signal (SV) but also the noise (LV), which implies a multiplicative additional number of samples are required.

## 3.3 Methods

### 3.3.1 How many samples

First assume that the populations to be sampled are finite and N=2000 and that their values are randomly distributed in space so that it is not necessary to consider the issue of "where to sample" in this section, but only the issue of "how many samples". In STATGRAPHICS, normal and lognormal populations were simulated with a user-specified average and variance. Four populations were generated for each distribution type, all with a average of 10 and each with a variance, expressed as relative standard deviations (RSD) of 25%, 50%, 100% and

150% (Figure 3.1). In order to determine the minimum sample size (MSS) required to estimate the average and variance of these eight populations, two methods were employed.

| Distribution | | Average | RSD(%) | Skew | Kurt |
|---|---|---|---|---|---|
| 1 | Normal | 10 | 25 | -0.04 | 0.06 |
| 2 | | | 50 | 0.01 | 0.06 |
| 3 | | | 100 | -0.02 | 0.09 |
| 4 | | | 150 | 0.01 | -0.05 |
| 5 | Lognormal | 10 | 25 | 0.69 | 0.51 |
| 6 | | | 50 | 1.92 | 8.19 |
| 7 | | | 100 | 6.6 | 93.22 |
| 8 | | | 150 | 7.2 | 98.68 |

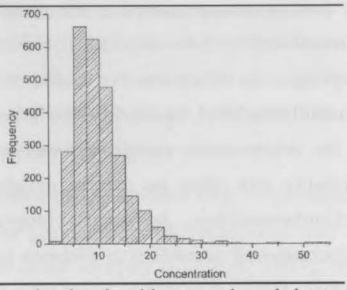



Figure 3.1 Skewness and kurtosis of eight populations simulated with normal and lognormal distributions, four relative standard deviations (RSD), average of 10 and N=2000. The graph presents the probability frequency distribution of one of the eight simulated populations: lognormal with RSD~50%, corresponding to a skewness of 1.92 and kurtosis of 8.19.

The first method is an iterative simulation procedure known as sampling-without-replacement (exemplified in Figure 3.2). One by one, samples were drawn at random and without replacement from each population, until the cumulative number of samples was enough to estimate the desired statistic (average and variance within a desired margin of error (ME= 20%, 15%, 10% and 5%). The MSS for a given ME was found by dividing the cumulative sample's estimate by the population's parameter, until the ratio fell somewhere inside the interval defined by each ME: 20%=0.80-1.20, 15%=0.85-1.15, 10%=0.90-1.10 and 5%=0.95-1.05 (exemplified in Figure 3.2). This process was repeated 2000 times (sampling rounds) to determine the MSS at four statistical significance levels (85%, 90%, 95% and 99%, two-tailed). Note that this procedure uses the sampling strategy known as simple-random sampling (see results section "Where to sample?").

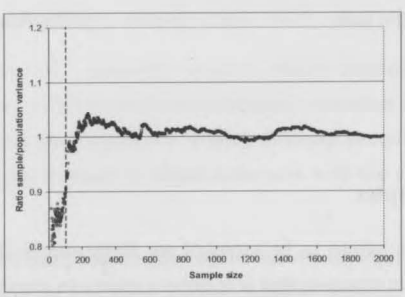



Figure 3.2 Example of the application of the simulation procedure, sampling-without-replacement, to the determination of the minimum sample size (MSS). In this sampling round, a sample size of n=103 (vertical slashed line) is enough to reach a margin of error of 10% (corresponding to the interval 0.90-1.10 on the y-axis) of the population's variance. This procedure was repeated 2000 times (sampling rounds) in order to determine the MSS required for different significance levels.

The second method is well-established sample-size formulas for finite populations (Table 3.4). Sample-size formulas exist that calculate the MSS required to estimate the average of normal and lognormal finite populations (Hale, 1972; StatPoint Technologies, Inc). To the best of our knowledge, formulas that calculate the MSS required to estimate the variance of finite populations are not available for any distribution. All formulas require the specification of a known or assumed population's size and variance, as well as the ME and significance level desired for the estimates. These four parameters were attributed the same values as those used in the sampling-without-replacement simulation described above (population size=2000; RSD=25%, 50%, 100% and 150%, ME=20%, 15%, 10% and 5%, significance level: 85%, 90%, 95% and 99%).

Table 3.4 Sample-size formulas that calculate the minimum sample size (MSS) required to estimate the average and variance of normal and lognormal populations (columns) when their size is finite and infinite (rows) (Cochran, 1977; Hale, 1972; StatPoint Technologies, Inc). To the best of our efforts, sample-size formulas to estimate the variance could only be found for the case of normal infinite populations (Cochran's theorem; Cochran, 1977; Hale, 1972; StatPoint Technologies Inc).

| | | Distribution | |
|---|---|---|---|
| Population size | Estimate | Normal | Lognormal |
| Finite (known) | Average | $\dfrac{\left(z_{\alpha/2}^2\sigma^2\dfrac{N}{N-1}\right)}{\left(\varepsilon^2+\dfrac{z_{\alpha/2}^2\sigma^2}{N-1}\right)}$ | $\dfrac{z_{\alpha/2}^2\sigma^2 N}{N\ln^2(\varepsilon+1)+z_{\alpha/2}^2\sigma^2}$ |
| | Var | Not found | Not found |
| Infinite (not known) | Average | $\left(\dfrac{z_{\alpha/2}\sigma}{\varepsilon}\right)^2$ | $\dfrac{z_{\alpha/2}^2\sigma^2}{\ln^2(\varepsilon+1)}$ |
| | Var | $\sigma\left(1-\sqrt{\dfrac{(n-1)}{\chi^2_{(\alpha/2),(n-1)}}}\right)\le\varepsilon$ and $\sigma\left(\sqrt{\dfrac{(n-1)}{\chi^2_{1-(\alpha/2),(n-1)}}}-1\right)\le\varepsilon$ (Cochran's theorem) | Not found |

$\sigma$-population's standard deviation (known or assumed); N-population's size if it is assumed or known to be finite; $\varepsilon$-desired margin of error; $z_{\alpha/2}$-desired two-tailed significance level for the normal distribution; $\chi^2_{(\alpha/2),(n-1)}$-is the desired two-tailed significance level for the chi-squared distribution.

The two methods above assumed finite populations with N=2000. However, sample-size formulas also exist to calculate the MSS required to estimate the average and the variance of infinite populations (Table 3.4). In this case, and to the best of our knowledge, formulas are available only for the normal distribution (Hale, 1972; StatPoint Technologies, Inc). Again these formulas require the specification of the known or assumed population's variance as well as the desired margin of error and statistical significance. These parameters were

attributed the same values as those in the sampling-without-replacement simulation described above.

A complementary investigation was carried out in order to determine whether the population's skewness/kurtosis affected the MSS. To this end, we resorted to real data, kindly provided by the Slovenian lichen survey (Jeran et al, 1996, 2003; reproduced with permission). The data concerns two chemical elements Calcium (Ca) and Cobalt (Co) selected precisely for having identical RSD (~52%) but very different skewness and kurtosis (and average) (Figure 3.8). There were three missing observations for Co, which were excluded from Ca, yielding a sample size of 187. The Lilliefors test statistic suggests that neither element can be considered normally distributed (not shown). The sampling-without-replacement procedure, described above, was performed on both data, which in this exercise function as populations. Owing to the small number of observations available (187), the margin of error was set at 20%. The simulation was repeated 100 times (sampling rounds) to calculate the MSS at each of four significance levels (85%, 90%, 95% and 99%, two-tailed).

### 3.3.2 Simulation of a population with spatial structure

A finite population of 10 000 observation, which can be taken to represent concentrations of some substance of interest, was simulated on a surface. Over the surface, the location of the observations is random, but the location of their values (concentrations) is not entirely random, having been purposely simulated so as to have two areas of exceptionally high values (hotspots) (Figure 3.3).

| | |
|---|---|
| Average | 14.14 |
| SD | 14.11 |
| RSD(%) | 99.79 |
| Skew | 2.40 |
| Kurt | 6.65 |
| Med | 8.41 |
| Min | 3.04 |
| Max hotspot 1 | 104.80 |
| Max hotspot 2 | 103.90 |
| P10 | 4.33 |
| P25 | 5.30 |
| P75 | 17.23 |
| P90 | 32.24 |

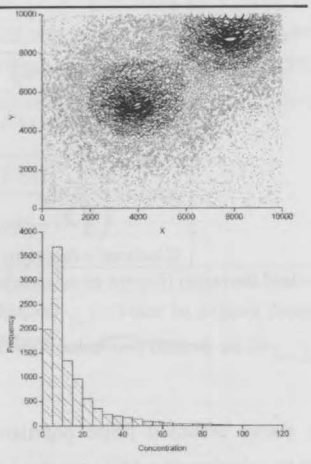



Figure 3.3 The simulated population (N=10 000) represented on a surface, its frequency distribution and its descriptive statistics. The size of the circles can be taken to represent the concentration of some substance of interest, whereas the hotspots can be taken to represent two point-sources for the substance. RSD-relative standard deviation; P10-P90-percentiles.

In STATGRAPHCS, this population was generated by first choosing a random location for the two hotspots. A concentration of 100 (arbitrary units) was attributed to each hotspot and concentrations were made to decrease exponentially with distance from them in all directions (isotropic). Finally, a random background variability of 3-5% was added to all observations. After a certain distance from the hotspots, the concentrations are completely randomly distributed. The resulting population has an average of 14, RSD of 100%, skewness of 2.4 and kurtosis of 6.7 (Figure 3.3).

This population was used in all remainder investigations, described below.

### 3.3.3 Where to sample?

The aim is to compare the performance of commonly used non-stratified probability-based sampling methods in their ability to accurately and precisely estimate several parameters of the population simulated above (methods section "Simulation of a population with spatial structure", Figure 3.3), so the number of samples was kept fixed. The number of samples should be of no influence to the comparison, so long as it is the same for all sampling methods. Preliminary investigations indicated that 400 samples should give a reasonable compromise between representativity and feasibility (not shown). Consultation of the Appendix (skewness~2 and kurtosis~7) suggests that 400 is an appropriate number of samples to estimate this population's variance at a 15% margin of error and 95% statistical significance level.

Four hundred samples were drawn at locations dictated by three non-stratified probability-based sampling methods: simple-random (R), systematic-grid (G) and random-within-blocks (B) (Figure 3.4 and Figure 3.5; and see their description in results section "Where to sample") (Chaudhuri & Stenger, 2005; Cochran, 1977). The last two methods require the placement of a sampling-grid of 20x20, where each grid-division defines a sampling site (also known as block) from which a single sample is drawn. For this exercise, the samples are drawn from within the sampling sites, but they might as well be taken at grid-nodes. Furthermore, the sampling sites are square but they could have virtually any shape. It is likely that the shape of the sampling sites is of some influence to estimation, but this issue will not be addressed here.

To evaluate the accuracy and vulnerability to sampling variability (i.e. precision) of the three sampling methods, 30 sampling surveys were performed, each time drawing 400 samples with the three sampling methods. Estimates of the population's average, variance, skewness and kurtosis obtained from each sampling method and survey were compared to the corresponding population's parameter with a ratio. For a closer inspection of the three sampling methods' ability to estimate the population's maximum values, corresponding to the two hotspots, the number of surveys was raised to 100.

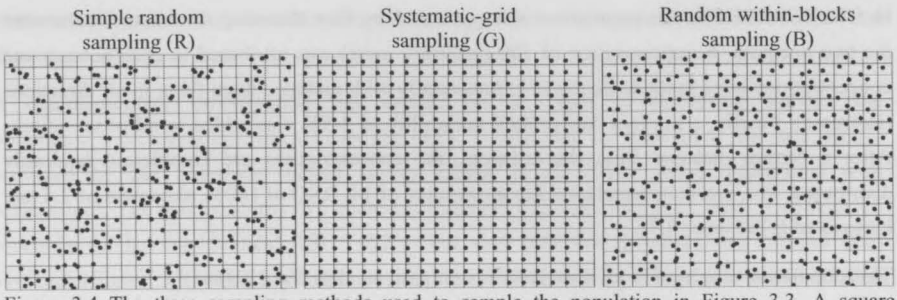



Figure 3.4 The three sampling methods used to sample the population in Figure 3.3. A square sampling-grid of 20x20=400 sampling sites or blocks was used.

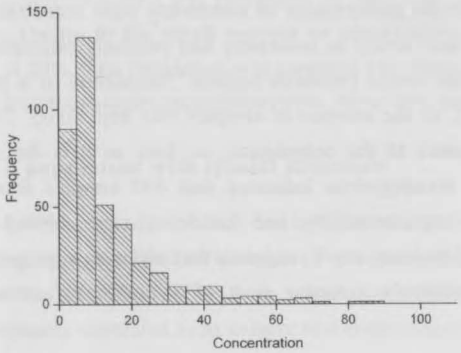



Figure 3.5 Frequency distribution of 400 samples taken from a 20x20 grid (400 blocks) from the population (N=10 000) in Figure 3.3. Each sample represents the arithmetic average of all samples within each block (composite sampling). The averages from each of the 400 blocks are plotted in the histogram.

### 3.3.4 The effect of the survey's sampling density

The aim is to evaluate the impact of the survey's sampling density on the estimates of the population described above (methods section "Simulation of a population with spatial structure", Figure 3.3). This was performed in the optimal situation where all observations in the population are sampled but their aggregation differs depending on the survey's sampling density (i.e. the number of sampling sites). In this case, where more than one sample is drawn from each sampling site, the survey's sampling density defines the number of sampling sites, and the multiple samples drawn from each sampling site are referred to as sub-samples. The sub-samples are pooled to form a composite sample which is then used to represent each sampling site.

Three survey sampling densities were compared: 20x20, 10x20, and 10x10, corresponding to 400, 200 and 100 sampling sites, respectively. All observations falling within each sampling site defined by the survey's sampling density were drawn (sub-samples) and pooled to form a composite sample. Each composite sample's average was calculated and used to represent each sampling site. Survey's estimates were then calculated from the values of the composite

samples resulting from each of the three survey's sampling densities. Twenty surveys were performed for each of the three sampling densities. Random-within-blocks (B) sampling was used to place the sampling-grid, since this was deemed the best feasible sampling method (see results section "Where to sample?").

Estimates of the population's average, variance, skewness, kurtosis, and maxima obtained from each of the three survey's sampling densities were calculated and averaged over the 20 surveys and finally compared to the corresponding population's parameters.

### 3.3.5 The effect of the sampling site's sampling density

The aim is to evaluate the impact of the sampling sites' sampling densities on the estimates of the population described above (see methods section "Simulation of a population with spatial structure", Figure 3.3), and on the estimates of the local populations defined by the survey's sampling density. This was investigated by using a fixed survey sampling density, and varying the number of sub-samples drawn from each sampling site. Random-within-blocks (B) sampling was used to place the sampling-grid and draw the sub-samples, since this was deemed the best feasible sampling method (see results section "Where to sample?").

Twenty surveys were performed, each time drawing 12, 10, 8, 6 or 4 sub-samples from each of the 400 sampling sites. The sub-samples within each sampling site were pooled to form a composite sample and their average was calculated. Finally, estimates of the population's average, variance, skewness, kurtosis and maxima as well as estimates of the sampling sites' (i.e. local populations) average and variance were calculated and compared to the respective parameters in the population (Figure 3.3).

## 3.4 Results

### 3.4.1 How many samples?

In this section, the minimum sample size (MSS) required to estimate the average and the variance of populations with diverse characteristics are determined.

The eight populations were simulated so as to mimic the characteristics of populations commonly encountered in, but not restricted to, sampling surveys of environmental parameters (Tables 3.1-3.3). These characteristics, some of which might be determinant for MSS requirements, included: normal and lognormal distributions with relative standard deviations (RSD) of 25%, 50%, 100% and 150%, skewness ranging from 0 to 7 and kurtosis ranging from 0 to 99 (Figure 3.1 shows the frequency distribution of one such population as well as some relevant descriptive statistics for all eight populations).

Three methods were used to calculate the MSS required to estimate each statistic (average and variance) in each population: 1) an iterative simulation procedure known as sampling-without-replacement (exemplified in Figure 3.2), 2) well-established sample-size formulas for finite populations (Table 3.4) and 3) well-established sample-size formulas for infinite populations (Table 3.4) (Hale, 1972; StatPoint Technologies, Inc).

The appendix shows the complete numerical results of the MSS required to estimate the average and variance, as calculated by each of the three methods, in the eight populations and at all margins of error (20%, 15%, 10% and 5%) and statistical significance levels (85%, 90%, 95% and 99%) being considered. In this section, we shall contrast only a part of those results graphically in order to deduce overall patterns.

According to the sampling-without-replacement simulation, the MSS required to estimate the average increases as the population's RSD increases for both distributions. The same MSS is required for both normal and lognormal distributions with similar RSD (Figure 3.6 and Appendix). The MSS obtained from sample-size formulas for finite populations display the same pattern but are somewhat lower, by about 20% on the average (Figure 3.6 and Appendix). The MSS obtained from sample-size formulas for infinite populations are similar to the MSS obtained from sample-size formulas for finite populations; only when the population's RSD is large and the ME and statistical significance level are strict does the former become substantially larger than the latter (Table 3.5 and Appendix).

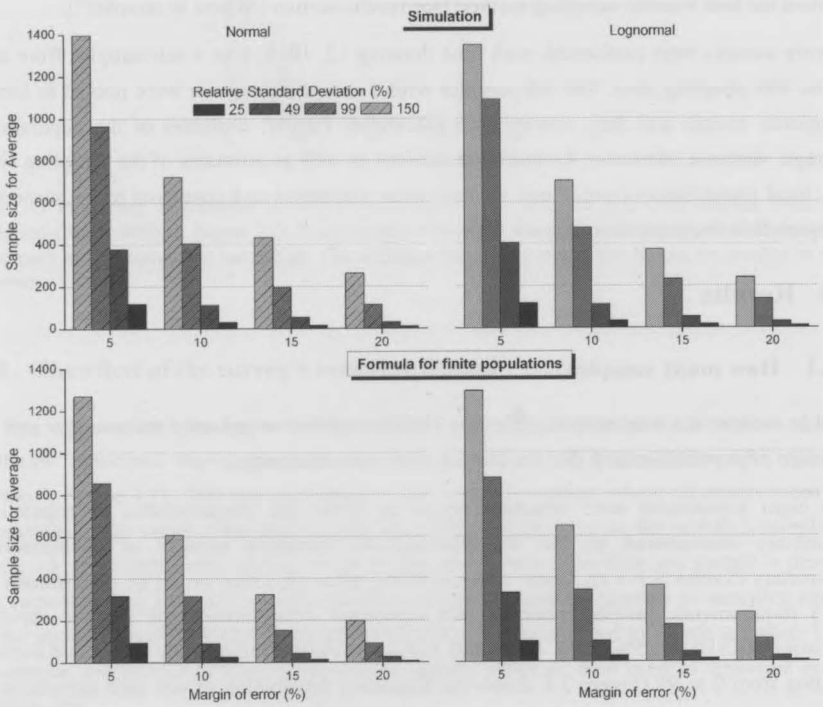



Figure 3.6 Minimum sample size (MSS, y-axes) required to estimate the average of normal and lognormal finite populations (N=2000) at the 95% statistical significance level, for four population's relative standard deviations (legend) and four margins of error (x-axes). MSS were calculated with a sample-size formula (right, Table 3.4) and with the sampling-without-replacement simulation procedure (left, Figure 3.2). The Appendix provides the complete numerical results and for all significance levels.

Table 3.5 Ratio between the minimum sample size (MSS) required to estimate the average of finite populations (N=2000) and that required to estimate the average of infinite populations. In both cases, the MSS were calculated with the appropriate sample-size formulas (see Table 3.4). The MSS required to estimate the average is the same regardless of the type of distribution (see Figure 3.6), therefore the ratios shown apply to both distributions. The Appendix provides the complete numerical results.

| | | ME | | | |
|---|---|---|---|---|---|
| Sign(%) | RSD(%) | 20% | 15% | 10% | 5% |
| 85 | 25 | 1.0 | 1.0 | 1.0 | 1.0 |
| | 50 | 1.0 | 1.0 | 1.0 | 0.9 |
| | 100 | 1.0 | 1.0 | 0.9 | 0.7 |
| | 150 | 0.9 | 0.9 | 0.8 | 0.5 |
| 90 | 25 | 1.0 | 1.0 | 1.0 | 1.0 |
| | 50 | 1.0 | 1.0 | 1.0 | 0.9 |
| | 100 | 1.0 | 0.9 | 0.9 | 0.7 |
| | 150 | 0.9 | 0.9 | 0.8 | 0.4 |
| 95 | 25 | 1.0 | 1.0 | 1.0 | 1.0 |
| | 50 | 1.0 | 1.0 | 1.0 | 0.8 |
| | 100 | 1.0 | 0.9 | 0.8 | 0.6 |
| | 150 | 0.9 | 0.8 | 0.7 | 0.4 |
| 99 | 25 | 1.0 | 1.0 | 1.0 | 0.9 |
| | 50 | 1.0 | 1.0 | 0.9 | 0.8 |
| | 100 | 0.9 | 0.9 | 0.8 | 0.4 |
| | 150 | 0.8 | 0.8 | 0.6 | 0.2 |

RSD- relative standard deviation; Sign-statistical significance; ME- margin of error.

According to the sampling-without-replacement simulation, the MSS required to estimate the variance is independent of the population's RSD in the case of normal distributions, whereas for lognormal distributions it increases as the population's RSD increases (Figure 3.7 and Appendix). To the best of our knowledge, the only sample-size formula available to estimate the variance applies only to normal infinite populations (Cochran's theorem; Cochran, 1977; Hale, 1972; StatPoint Technologies Inc) (Table 3.4). The MSS required to estimate the variance of normal finite populations obtained by simulation agree quite well with that obtained from Cochran's theorem, except when the margin of error and statistical significance are very strict, in which case Cochran's theorem recommends substantially larger MSS than the simulation procedure, likely because the former is meant for infinite populations whereas the latter applies to finite N=2000 populations (Table 3.6 and Appendix).

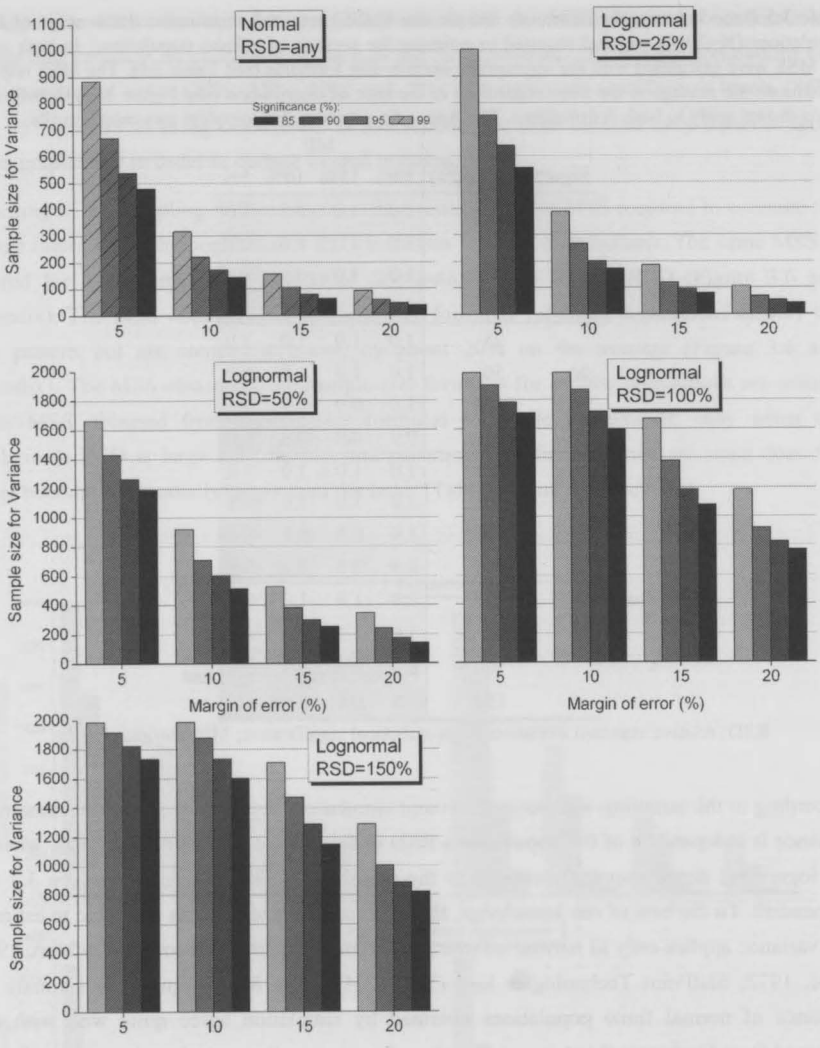



Figure 3.7 Minimum sample size (MSS) required to estimate the variance of normal and lognormal finite populations (N=2000) for four statistical significance levels (legend) and four margins of error (x-axes) and, in the case of lognormal populations, for four relative standard deviations (RSD) (individual graphs). MSS were calculated with the sampling-without-replacement simulation (see Figure 3.2). Note that, in the case of normal populations, the MSS is independent of the RSD (legend: RSD=any). The appendix provides the complete numerical results.

Table 3.6 Ratio between the minimum sample size (MSS) required to estimate the variance of finite normal populations (N=2 000) and that required to estimate the variance of infinite normal populations. The MSS for finite populations was calculated with the sampling-without-replacement simulation, whereas that for infinite normal populations was calculated with Cochran's theorem (see Table 3.4). The appendix provides the complete numerical results.

| | | ME | | | |
|---|---|---|---|---|---|
| Sign(%) | RSD(%) | 20% | 15% | 10% | 5% |
| 85 | Any | 1.0 | 1.0 | 1.1 | 1.0 |
| 90 | | 0.9 | 1.0 | 1.0 | 0.9 |
| 95 | | 0.8 | 0.9 | 0.9 | 0.8 |
| 99 | | 0.8 | 0.8 | 0.8 | 0.6 |

Sign-statistical significance; ME-margin of error.

The fact that in normal populations (where skewness and kurtosis are zero independently of RSD), the MSS required to estimate the variance is independent of the RSD, whereas in lognormal populations (where skewness and kurtosis increase as the RSD increases), the MSS required to estimate variance increases as the RSD increases, suggests that skewness and kurtosis, rather than RSD, are the determinant factors for the MSS required to estimate the variance.

| | Ca | Co |
|---|---|---|
| N | 187 | 187 |
| Mean | 22768 | 0.38 |
| SD | 11786 | 0.20 |
| Skewness | 0.83 | 2.61 |
| Kurtosis | 1.07 | 9.36 |
| RSD | 51.77 | 51.58 |

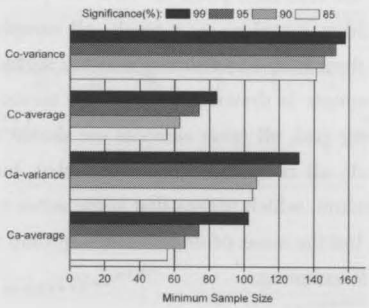



Figure 3.8 Table: Descriptive statistics for Calcium (Ca) and Cobalt (Co) concentrations measured in the Slovenian lichen survey (Jeran et al, 2003; reproduced with permission). Graph: minimum sample size required to estimate the average and variance of Ca and Co, within a margin of error of 20% and for four statistical significance levels (legend). The MSS were calculated with sampling-without-replacement simulation.

To confirm this hypothesis, we resorted to real data kindly provided by the Slovenian lichen survey (Jeran et al, 2003; reproduced with permission). The data pertains to two chemical elements: Calcium (Ca) and Cobalt (Co) selected precisely for having a similar RSD (~52%) and very different skewness and kurtosis (Figure 3.8). Sampling-without-replacement from the two elements suggests that, in this particular data and for a 20% ME, 30% more samples are required to estimate the variance of the more skewed/kurtosic element, whereas the same number of samples is required to estimate the average of both elements (Figure 3.8). Since the only population's parameter considered in available sample-size formulas is the population's variance, it is perhaps no wonder that no sample-size formulas exist (to the best of our

knowledge) to estimate the variance of lognormal populations; according to the above observations, formulas would need to consider the population's skewness and kurtosis instead of the population's variance.

Obviously, larger MSS are required to estimate the average than the variance for a given set of conditions (RSD, skewness/kurtosis, ME and statistical significance level) and larger MSS are required as the ME and statistical significance level become stricter.

### 3.4.2 Where to sample?

The question of where to sample is essentially a question of choosing a sampling method of the many available (e.g.: Chaudhuri & Stenger, 2005). Here we compare just three of the most common non-stratified probability-based sampling methods,; they differ with respect to how regular or random the samples are drawn from the surface.

The three sampling methods are illustrated in Figure 3.4. With simple-random sampling (R), each and every sample is drawn randomly from anywhere on the surface (irrespective of any sampling-grid); as a result, some areas of the surface are sampled with less density than others, and the size of the areas of under-sampling (and thus of over-sampling) can be very variable across the surface. With systematic-grid sampling (G), the first sample is drawn randomly and serves as the reference point for both the placement of the sampling-grid and for the location of all other samples; as a result, all sampling sites are sampled, and all samples are equidistant from each other throughout the surface. With random-within-blocks sampling (B), the first sample is drawn randomly and serves as the reference point for the placement of the sampling-grid, all other samples are drawn at random locations within the sampling sites; as a result, all sampling sites are sampled, but the samples' location within each sampling site is random, which means that some areas of the surface are sampled with less density than others, but the areas of under-sampling (and thus of over-sampling) are less wide than in the case of R-sampling.

Figure 3.3 shows the population (N=10 000, positively skewed with RSD=100%) to be sampled. The circles' size can be taken to represent the concentration of some substance of interest in a discrete medium amenable to being sampled. The concentrations are not entirely randomly distributed over the surface since high values tend to cluster around two hotspots which could be taken to represent point-sources such as volcanoes, industrial facilities or laundry shops. The concentrations decrease exponentially and isotropically with distance from the hotspots and after a certain distance, the concentrations are entirely randomly distributed.

The aim is to compare the three sampling methods with regards to how accurately and precisely they are able to estimate the average, variance, skewness, kurtosis and maxima of the population described in the previous paragraph. To this end the survey's sampling density was kept constant at 400 samples (i.e. in the case of G- and B-sampling this corresponds to 400 sampling sites defined by the sampling-grid and one sub-sample is drawn from each of the 400 sampling sites). The estimates obtained from each sampling method were compared

to the corresponding population's parameter by means of a ratio, which quantifies the margin of error of the estimates. Multiple surveys were carried out in order to investigate the susceptibility of the estimates to sampling variability, depending on the sampling method used.

With regards to the estimation of the average, variance, skewness and kurtosis, on the average (over 30 surveys), G- and B-sampling performed similarly better in that the sampling variability around the true population's parameter was lower compared to that of R-sampling (Figure 3.9). All three sampling methods showed greater sampling variability in estimating higher moments (kurtosis and skewness) than the average and variance, but nevertheless all three methods were fairly accurate (i.e. centred on the true parameter).

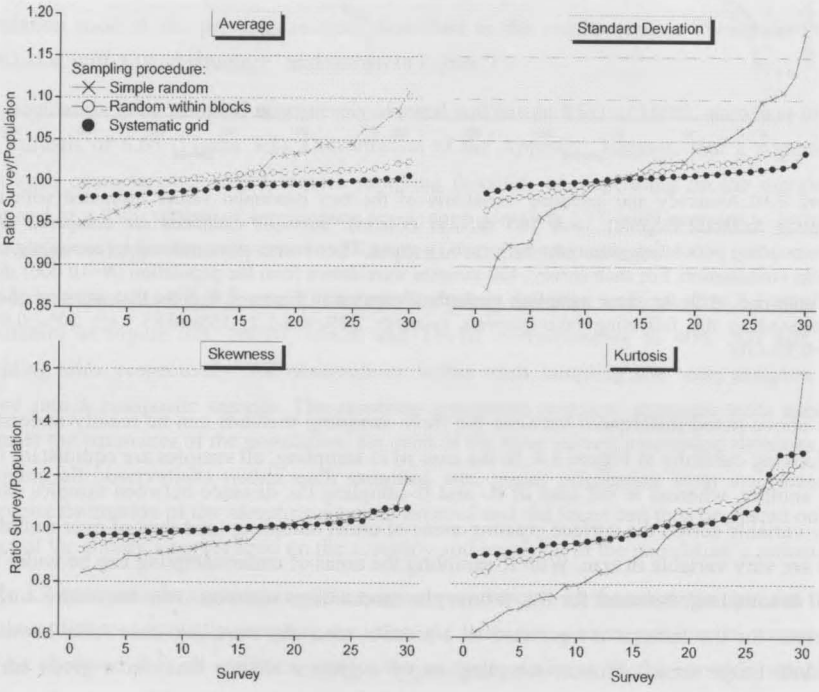



Figure 3.9 Accuracy and sampling variability of the average, variance, skewness and kurtosis estimated with three sampling methods (legend), over 30 surveys (x-axes). Survey's estimates are compared to the corresponding population parameter by a ratio (y-axes). The surveys were ordered by ascending ratios to help visualisation. For each survey, 400 samples were drawn from the population (N=10 000) shown in Figure 3.3, with the three sampling methods illustrated in Figure 3.4. Note that margins of error correspond to the following ratio-intervals (y-axes): 20%=0.8-1.2; 15%=0.85-1.15; 10%=0.9-1.1, 5%=0.95-1.05.

With regards to the estimate of the two maximum values in the population, corresponding to the two hotspots, on the average (over 100 surveys), G-sampling performed considerably better than R- and B-sampling since the former's estimates of the maxima showed less variability around the true population's maxima compared to the other two sampling methods (Figure 3.10).

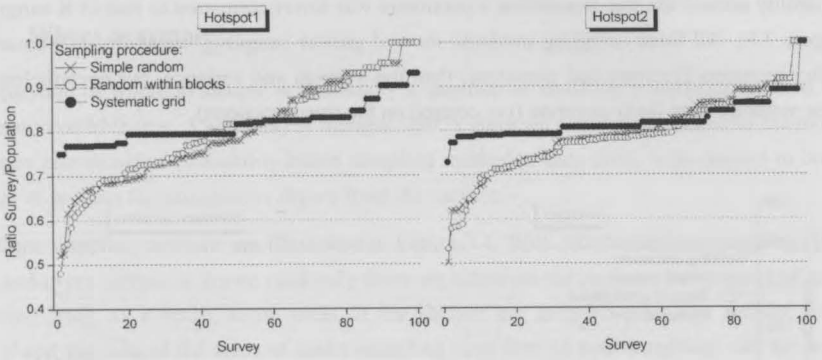



Figure 3.10 Accuracy and sampling variability of the two maximum values estimated with three sampling methods (legend), over 100 surveys (x-axes). Survey's estimates are compared to the corresponding population parameter by a ratio (y-axes). The surveys were ordered by ascending ratios to help visualisation. For each survey, 400 samples were drawn from the population (N=10 000) shown in Figure 3.3, with the three sampling methods illustrated in Figure 3.4. Note that margins of error correspond to the following ratio-intervals (y-axes): 20%=0.8-1.2; 15%=0.85-1.15; 10%=0.9-1.1, 5%=0.95-1.05.

The above-noted differences between the three sampling methods can be readily understood by looking carefully at Figure 3.4. In the case of G-sampling, all samples are equidistant from one another, whereas in the case of R- and B-sampling the distance between samples can be very variable across the surface creating areas of under-sampling (and thus of over-sampling) that are very variable in size. With R-sampling the areas of under-sampling can be wider than with B-sampling, because for the former the randomness operates over the entire surface, whereas for the latter it is restricted to a smaller area, the sampling sites (also known as blocks). Large areas of under-sampling imply a greater chance that, on a given survey, hotspots will not be sampled, which of course affects more greatly the estimates of the maxima and consequently will also trim down the estimates of kurtosis, skewness and finally of the RSD and average.

Notwithstanding the general conclusion above that, for a given survey's sampling density, G-sampling affords greater precision for a wider range of statistics than the other two sampling methods tested, in practice, the researcher will often tolerate estimates within a certain margin of error (ME). Noting that ME corresponds to the following intervals on the y-axes of Figure 3.9 and Figure 3.10: 20%=0.8-1.2; 15%=0.85-1.15; 10%=0.9-1.1, 5%=0.95-1.05, it becomes clear that, unless the chosen ME is very strict, the three sampling methods may be considered identically precise at estimating the population's average and variance. With respect to the

population's maxima, G-sampling is prominently more precise than the other two methods even at a large ME (20%). Considering that the maxima might be the most interesting and useful feature of such a population (Figure 3.3), G-sampling is the most effective sampling method. However, G-sampling can rarely be executed with fidelity in the field because samples can rarely be found or placed at the exact planned coordinates. Therefore, all things considered, B-sampling appears to offer the best compromise between feasibility and precision in estimation for a wide range of statistics, with the exception of tail-values.

### 3.4.3 1.1.1 The effect of the survey's sampling density

The survey's sampling density defines the number (and size) of the sampling sites and thus the level of aggregation of the original population or, in still other words, the resolution of the survey. Here we investigate how the survey's sampling density impacts the estimation the population used in the previous section, described in the results section "Simulation of a population with spatial structure" and shown in Figure 3.3.

The population to be sampled is positively-skewed and has an RSD of 100%, skewness of 2.4 and kurtosis of 6.65 (Figure 3.3). Consultation of the Appendix suggests that a population with such characteristics might require sampling densities of, depending on the margin of error and at a 95% statistical significance level, approximately 137-1089 samples to estimate the average and approximately 240-1432 samples to estimate the variance.

Three, less than ideal but practical, survey's sampling densities were used to estimate the population in Figure 3.3: 20x20, 10x20 and 10x10, corresponding to 400, 200 and 100 sampling sites, respectively. All observations within each sampling site were sampled and polled into a composite sample. The resulting composite samples' averages were used to estimate the parameter of the population, for each of the three survey's sampling densities. By sampling all observations within each sampling site, issues originating from inaccurate or imprecise estimation of the sampling sites are avoided and the focus can thus be placed on the effect of the survey's aggregation on the accuracy and precision of the population's estimates.

Table 3.7 compares the parameters of the original population with the estimates obtained from the three different survey's sampling densities; the latter being averaged over 20 surveys. It can be observed that all statistics, except the average, decrease as the survey's sampling density decreases. The decrease is particularly pronounced for the maxima and minima, followed by the skewness and kurtosis and finally for the RSD.

Figure 3.11 shows the concentrations as observed in the original population and as estimated with the survey's sampling density of 400 sampling sites, along x- and y-cross-sections over the two hotspots. It is easily observed that all values are smoothed by the aggregation: high values become lower and low values become higher and that the transition between observations/sampling sites is smoother, losing the fine irregular detail that is present in the original disaggregated population.

Table 3.7 Comparison of the true parameters of the population (first column, see Figure 3.3) with the estimates obtained from three survey sampling densities: 20x20, 10x20, 10x10; corresponding respectively to 400, 200 and 100 sampling sites. All observations within each sampling site were pooled. Estimates are averages of 20 surveys.

| | Sampling density | | | |
|---|---|---|---|---|
| | Population | 400 | 200 | 100 |
| Mean | 14.14 | 14.26 | 14.28 | 14.23 |
| SD | 14.11 | 14.07 | 13.97 | 13.50 |
| RSD(%) | 99.79 | 98.67 | 97.78 | 94.86 |
| Skew | 2.4 | 2.30 | 2.24 | 2.03 |
| Kurt | 6.65 | 5.92 | 5.54 | 4.04 |
| Median | 8.41 | 8.37 | 8.41 | 8.26 |
| Minimum | 3.04 | 3.93 | 4.01 | 4.01 |
| Max hotspot 1 | 104.8 | 87.55 | 80.90 | 67.24 |
| Max hotspot 2 | 103.9 | 83.91 | 77.77 | 62.18 |
| P10 | 4.33 | 4.40 | 4.36 | 4.49 |
| P25 | 5.3 | 5.22 | 5.17 | 5.30 |
| P75 | 17.23 | 17.82 | 17.65 | 18.14 |
| P90 | 32.24 | 31.05 | 30.84 | 31.54 |

RSD-relative standard deviation; P10-P90-percentiles.

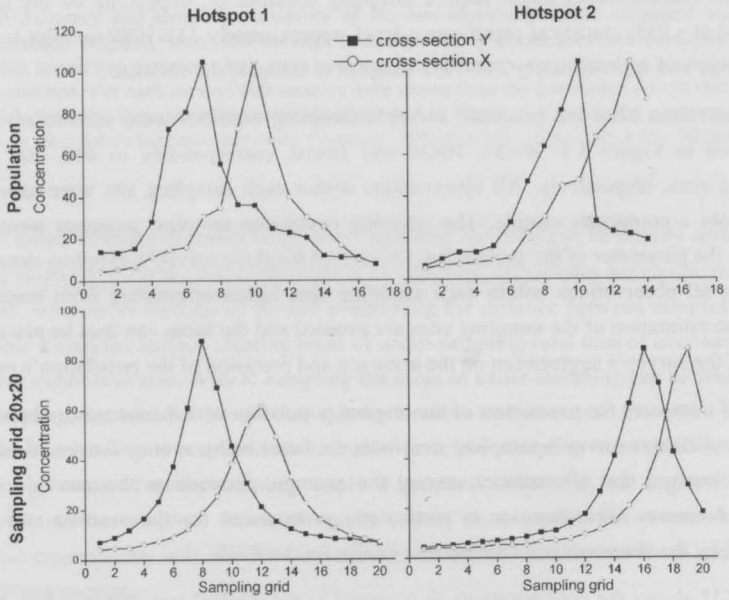



Figure 3.11 X- and Y-cross-section over the two hotspots. Top: values of the actual observations in the population shown in Figure 3.3. Bottom: estimates obtained from a 20x20 sampling grid where all observations within each of the 400 sampling sites are pooled and averaged.

### 3.4.4 The effect of the sampling sites' sampling density

In practice, it is never possible to sample all observations in the population as was done in the previous section, even if they are meant to be pooled and analysed as composite samples. In

most atmospheric biomonitoring surveys, for instance, only 5 sub-samples are drawn from each sampling site (Wolterbeek et al, 2010). While the survey's sampling density defines the number of sampling sites of the survey, the sampling site's sampling density defines the number of sub-samples drawn from each sampling site. In this section, the aim is to assess, for a fixed survey's sampling density of 400 sampling sites, the impact of drawing different numbers of sub-samples, on the estimates of the population as well as on the estimates of the local populations within the sampling sites. The population to be estimated is the same as in the previous two sections and is described in the results section "Simulation of a population with spatial structure" and shown in Figure 3.3.

For the population in Figure 3.3 and a survey's sampling density of 400 sampling sites, on average (i.e. over the 400 sampling sites and over 20 surveys), the local populations (i.e. all observations within sampling sites) have an RSD of about 10% (min-max: 4-18%). The primary aim of most surveys is to estimate the population's variance, but in order to do so one needs to accurately estimate the sampling sites' averages. Resorting to the sampling-without-replacement simulation procedure or to sample-size formulas (Figure 3.2 and Table 3.4), one easily finds that estimation of the average in populations with a 10% RSD requires 1-15 sub-samples depending on the desired margin of error and for a 95% significance level. Thus the number of sub-samples being tested here (12, 10, 8, 6 and 4) should give a reasonable range for the estimation of the local averages.

Table 3.8 Comparison of the effect of the number of sub-samples (n) taken from each of 400 sampling sites (20x20 sampling-grid) on the estimates of the population shown in Figure 3.3. The number of sub-samples within each sampling site varies from all observations to 12, 10, 8 or 4 sub-samples. Sub-samples were pooled to give the average per sampling site. The sampling procedure was repeated 20 times (surveys) and the results show the average and standard deviation over those surveys.

| | n=all | n=12 | | n=10 | | n=8 | | n=4 | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Mean | 14.26 | 14.26 | 0.03 | 14.25 | 0.03 | 14.26 | 0.04 | 14.25 | 0.06 |
| SD | 14.07 | 14.08 | 0.06 | 14.09 | 0.07 | 14.09 | 0.07 | 14.08 | 0.13 |
| RSD(%) | 98.67 | 98.74 | 0.31 | 98.83 | 0.34 | 98.81 | 0.35 | 98.80 | 0.67 |
| Skew | 2.30 | 2.31 | 0.02 | 2.31 | 0.03 | 2.31 | 0.03 | 2.31 | 0.04 |
| Kurt | 5.92 | 5.95 | 0.19 | 6.02 | 0.25 | 5.99 | 0.27 | 6.00 | 0.35 |
| Med | 8.37 | 8.37 | 0.09 | 8.37 | 0.09 | 8.37 | 0.09 | 8.36 | 0.14 |
| Min | 3.93 | 3.77 | 0.09 | 3.75 | 0.09 | 3.71 | 0.09 | 3.55 | 0.13 |
| Hotspot 1 | 87.55 | 87.21 | 2.25 | 87.75 | 3.02 | 87.27 | 3.18 | 87.19 | 4.62 |
| Hotspot 2 | 83.91 | 83.93 | 1.65 | 83.98 | 2.12 | 83.63 | 2.11 | 82.16 | 3.54 |
| P10 | 4.40 | 4.40 | 0.04 | 4.40 | 0.04 | 4.40 | 0.04 | 4.41 | 0.04 |
| P25 | 5.22 | 5.19 | 0.03 | 5.17 | 0.04 | 5.17 | 0.05 | 5.20 | 0.06 |
| P75 | 17.82 | 17.82 | 0.14 | 17.82 | 0.17 | 17.82 | 0.18 | 17.72 | 0.20 |
| P90 | 31.05 | 31.44 | 0.38 | 31.43 | 0.48 | 31.46 | 0.51 | 31.49 | 0.62 |

RSD-relative standard deviation; P10-P90-percentiles.

Table 3.8 shows that, on the average (over 20 surveys), the number of sub-samples within each sampling site does not affect the estimates of the population (average, variance, skewness, kurtosis, maxima, etc) to any noticeable extent when compared to the case of the previous section where all observations within each of the 400 sampling sites are pooled. However, decreasing the number of sub-samples does increase the sampling variability of all the survey's estimates, especially (in descending order) of the maxima, kurtosis and skewness.

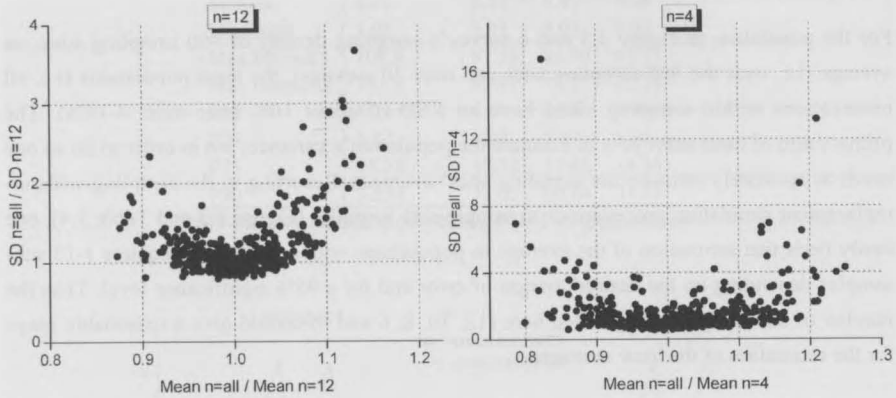



Figure 3.12 Mean and standard deviation (SD) of each of the 400 sampling sites obtained when each sampling site is represented by 12 sub-samples (left) and by 4 sub-samples (right). In both graphs the mean and SD were divided by the mean and SD obtained when all observations within each sampling site are sampled (n=all). Note that when n=4, estimates vary more greatly around one than when n=12.

Figure 3.12 shows, for one particular survey, the average and the variance of each of the 400 sampling sites (i.e. local averages and local variances) estimated with 4 and 12 sub-samples and compares them with the "true" local averages and "true" local variances obtained when all observations within each of the 400 sampling sites are sampled and pooled. It can be observed that the average of each sampling site across is unbiased but tends to be more distant from the "true" local average as the number of sub-samples decreases, denoting increasing vulnerability to sampling vulnerability. The local variance, on the other hand, tends to be systematically underestimated, and the degree of this underestimation again, tends to be greater as the number of sub-samples decreases. This should be expected as the number of sub-samples tested is far from sufficient to estimate the variance of even normal populations (Appendix). There is no indication that the bias in local variances is greater at sampling sites with greater "true" local variances (not shown).

## 3.5 Discussion & Conclusion

Minimum sample sizes (MSS) for estimating the average and variance have been laid out in the Appendix for normal and lognormal population with a range of RSD, skewness and kurtosis, inspired by, but certainly not restricted to, distributions observed in atmospheric biomonitoring surveys of airborne chemical elements (Table 3.1-3.3). The MSS provided apply to populations that are homogeneous, i.e. have no spatial/temporal structure.

To calculate the MSS required for estimating the average and variance of the populations in the appendix two methods were used: (1) a simulation procedure known as sampling-without-replacement, and (2) sample-size formulas for finite and infinite populations (Hale, 1972; StatPoint Technologies, Inc). Whenever comparisons between them were possible (depending on the availability of sample-size formulas for particular types of distributions) they revealed a good agreement. The sampling-without-replacement procedure, however, offers greater flexibility since it can calculate the MSS required to estimate any population statistics (e.g.: skewness, kurtosis) and not just the average and variance, it does not require that a distinction be made between normal and lognormal distributions (normality tests often diverge in their diagnostics, not shown), and it can be applied to populations with any distribution and not just normal and lognormal ones. This procedure does, however, require data (real or simulated) of a population with similar characteristics as the one to be sampled, and for this reason sample-size formulas, when available, are more practical to use.

The population characteristics that are determinant to the MSS required to estimate the average and variance were identified thanks to the sampling-without-replacement simulation. In the case of the estimation of the average, the determinant population characteristic is the RSD, independently of the type of distribution; whereas in the case of the estimation of the variance, the determinant population characteristic is the skewness and kurtosis, again independently of the population's distribution and indeed of its RSD. No sample-size formulas exist that calculate the MSS required to estimate the variance of lognormal populations, since all sample-size formulas use only the population's variance as input. The sampling-without-replacement procedure, however, can recommend MSS for estimating the variance of lognormal and other skewed distributions that are so commonly encountered in environmental data.

Three of the most commonly encountered probability-based sampling methods that differ in terms of the regularity/periodicity with which samples are drawn were applied to a population with spatial structure using a fixed survey's sampling density. All three methods gave unbiased estimates for all statistics (except maxima of course); however they were differently affected by sampling variability. Systematic-grid sampling appears to be the best sampling method of the three because its estimates are the least affected by sampling variability and this was true for all statistics estimated and, in particular, maximum values are considerably more precise with this sampling method than with the other two. Random-within-blocks sampling offers similar precision as systematic-grid sampling for all statistics except maxima, where sampling variability is substantially greater. The differences between the sampling

methods can be readily understood by noting the existence and the size of under-sampled areas relative to the resolution of the survey (i.e. the survey's sampling density). In simple-random sampling, under-sampled areas can be wide and if they fall in unusual areas such as hotspots the survey will fail to represent them. In the case of random-within-blocks, areas of under-sampling exist but are much smaller in size, and thus the chance that they will capture unusual areas such as hotspots, if only partially, is lower. Finally, with systematic-grid sampling, because all samples are equidistant from one another, no areas of under-sampling (or over-sampling) can be said to exist with respect to the survey's sampling density, so if unusual areas are not accurately estimated, it implies that the survey's sampling density is inappropriate rather than a problem of where the samples are drawn from. It follows that a survey carried out with, for instance, simple-random sampling will require a greater number of samples to estimate a given statistic with a similar precision as say, systematic-grid sampling. However, in practice, it is impossible to find or place samples at the exact coordinates planned by systematic-grid sampling, therefore any survey planned as systematic-grid sampling becomes, in the field, some form of random-within-blocks sampling. Most atmospheric biomonitoring surveys to date have been carried out in this way: planned as systematic-grid sampling and carried out as random-within-blocks (Jeran et al, 2003; Freitas et al, 1999; Ruhling, 1994). Thus, random-within-blocks sampling appears to be the best compromise between feasibility and estimation precision, with the least number of samples possible; however, if tail-values are the main interest of the sampling survey, systematic-grid sampling is considerably more effective than the other sampling methods. In populations with no spatial structure, however, the three sampling methods should provide identically precise estimates.

The survey's sampling density defines the level of aggregation of the population. From this perspective, it defines the amount and type of variability in the original disaggregated population that is lost to aggregation (i.e. resolution). Loss of information is desirable up to a point in order to remove noise and outliers, which affect disaggregated data more severely than averaged data. Loss of certain types of variability may also be indispensable for some purposes (e.g.: removal of seasonal fluctuations in daily time-series) (Cox, 2000; Lumley & Sheppard, 2000). In general, the sparser the sampling density the trimmer the distribution of the aggregated population becomes compared to the original disaggregated population. This is mostly mediated by the attenuation of maximum and/or minimum values (which may be desirable as tail-values could be outliers), which in turn trims, in descending order, the kurtosis, skewness and RSD. In some cases, depending on the population's distribution, changing the level of aggregation can also conceivably affect the estimate of the population's average considerably and/or even an entirely different distribution may arise (not shown). With regards to the estimates of the average of the sampling sites (i.e. local averages), a sparser survey's sampling density means that there is less contrast between them and the transition from one sampling site to another is smoother (which may be desirable up to a point as irregularities could be noise). The above is true whether local averages are calculated from the average of all individual sub-samples or from the values of partial or whole composite samples. This has implications for the planning of a sampling survey because in order to

calculate the MSS one must have an idea of the value of the population's parameters (e.g.: variance, skewness, kurtosis) and to obtain this information researchers often resort to previous surveys and small samples. From the discussion so far, it follows that the parameters provided from such sources are likely to be underestimated due to the high level of aggregation and/or reduced coverage of small samples and most surveys. This can lead to the calculation of an insufficient number of MSS, especially for estimating the population's variance, since the skewness and kurtosis of the population (remember skewness and kurtosis are the determinant factors for the MSS to estimate the variance) can be heavily smoothed (and imprecise) by too high an aggregation level.

The sampling site's sampling density, or equivalently, the number of sub-samples representing each sampling site defined by a given survey's sampling density was investigated in terms of its impact on both the survey's and the local population's estimates. It is a fundamental rule of nature captured by statistics that the lower the number of samples the greater the sampling variability in estimation. It was observed that sampling variability has greater impact the higher the moments being estimated (in descending order: maximum, kurtosis, skewness, RSD, average). Thus the sampling site's sampling density does not, on the average, bias the survey's estimates or the local averages, but it does affect their precision. The sampling site's sampling density does, however bias the local variances: local variances tend to be underestimated in the majority of sampling sites and this is all the more pronounced the lower the number of sub-samples. The under-estimation of local variances does not appear to be related to the true concentrations at the sampling sites (e.g.: greater under-estimation at sampling sites with true large concentrations).

Since population estimates (e.g.: survey's variance) are based on the values of the sampling sites defined by the survey's sampling density, the sampling sites need to be accurately and precisely estimated. The averages of the sampling sites (i.e. local averages) need to be accurately estimated because they are the building blocks for the calculation of all the survey's estimates. The variance within sampling sites (i.e. local variance) need to be accurately estimated because they provide the uncertainty of the survey's estimates: if the variance within the sampling sites is large, the uncertainty in the survey's estimates is correspondingly large. Thus it is necessary to have as rigorous an estimate of the population defined by the survey's sampling density as it is of the local populations within the survey's sampling sites. The examples that follow will hopefully illustrate this point and make the connection with the signal-to-noise ratio as expressed by the SV/LV-ratio (Wolterbeek et al, 1996; Wolterbeek & Verburg, 2002), as well as demonstrate the use of the Appendix in planning a sampling survey.

As a first example suppose we wish to estimate the variance of a population characterised by skewness of 2 and kurtosis of 8 (corresponding to a lognormal with RSD=50%) and we wish to do so at a statistical significance level of 95%. Consultation of the appendix recommends an MSS ranging from 240-1432 depending on the margin of error desired. It would be inefficient and misleading to demand a ME for the survey's estimates that is lower than the uncertainty contained within the survey's sampling units, i.e. the local variance. Thus assume

that the local populations have a known variance of 25%, then the ME for the survey's estimates might be set at no less than 25% (but we shall use a ME of 20% to follow the values given in the appendix), in which case 240 samples would be needed in order to estimate the population's variance. These 240 samples actually represent sampling sites for which it is necessary to accurately and precisely estimate the local averages in order to have an accurate and precise estimate of the population's variance. Thus, local populations with an RSD=25% estimated with, say a ME=10% (the margin of error for the local averages may be no less than for instance the analytical error) and significance level of 95%, would require 30 sub-samples per sampling site (Appendix). Overall, a population and survey with these characteristics would require 240x30 samples (i.e. 240 sampling sites each with 30 sub-samples).

In a second example, suppose that everything was identical to the first example, except that the local variance was known to be 10% instead of 25%. In this case, one might set the survey's ME at no less than 10%. In this case, one would require 709x5 samples (not shown in Appendix, calculated by formula in Table 3.4). Note that, compared to the first example, fewer sub-samples are required to estimate the local averages but more sampling sites are required to estimate the survey's population, reflecting a decrease in the uncertainty of the survey's estimates and thus an increased ability to capture more details at the survey level.

Consider now a third and final example, departing from the first. Suppose that an uncertainty of 25% for the survey's estimates, corresponding to the local variance in the first example, is deemed too high for the purposes of the research. The most obvious straightforward solution is to increase the survey's sampling density, i.e. to increase the number of sampling sites and thus decrease their size. The desired effect is that sampling sites should become more homogeneous, i.e. lower variance. The collateral effect is that decreasing the aggregation of the population (relative to the first example) can increase the population's kurtosis, skewness and RSD considerably. So suppose in a somewhat exaggerated example (to follow the values in Appendix) that the sampling sites were made smaller and as a consequence the aggregated population's skewness and kurtosis increased from 2 and 8 (first example) to 7 and 93, respectively. A population with these characteristics would now require 921-1902 sampling sites (Appendix) to estimate its variance, depending on the local variance (margin of error) achieved by the new survey's sampling density and for a 95% statistical significance level. Suppose that the local variances were known to decrease from 25% (first example) to 10%, the survey would then require 1872 sampling sites to estimate the population's variance and 5 sub-samples to estimate the local averages. This example hopefully makes it clear that decreasing the aggregation of the population can unveil more details and complexity which on its own, independently of the margin of error, requires a greater number of sampling sites; whereas in the second example there is simply a decrease in the uncertainty (margin of error) for a particular aggregation level of the population, which also demands a greater number of sampling sites. In practice, a decrease in the uncertainty and a decrease in the aggregation level of the survey go hand in hand.

If the investigator should decide to use composite sampling, the number of (composite) samples that would need to be analysed individually in the first, second and third examples

are 240, 790 and 1872 whereas the number of sub-samples forming the composite samples required to estimate the local averages would be 30, 5, and 5, respectively. Assuming that the local variances in the three examples were known with exactitude, the SV/LV ratio in the three examples would 50/25=2, 50/10=5 and 100/10=10, respectively.

The above examples were somewhat contrived for at least two reasons.

The first reason is that they considered the local variance as given independently of the survey's sampling density. In reality, local variances can only be defined with respect to the survey's sampling density. Changing the survey's sampling density changes the local variance and a change in local variances changes the margin of error possible for the survey's sampling density. Thus both local variance and the survey's sampling density must be considered in concert in order to decide on a sound number of sampling sites and sub-samples. Since the local variance defines the uncertainty in the survey's estimates, they need to be kept below a maximum level chosen to suit the purposes of the research. To achieve this, the researcher would probably need to perform several preliminary surveys in order to adjust the survey's sampling density (i.e. the size of the sampling sites) to the desired local variance. Once a suitable survey's sampling density is found that keeps local variances below a maximum level, it defines the degree to which the original population is to be aggregated. The degree of aggregation, in turn, defines the degree of detail remaining from the original disaggregated population; the properties of this aggregated population in turn need to be considered in defining a suitable number of sampling sites to estimate it, which again feedbacks on the local variance obtained from that number of sampling sites.

The second reason why the examples above were somewhat contrived is that they assumed an a priori known local variance, when in reality it needs to be measured. Assuming the best case scenario, where the local populations of the three examples are normally distributed (i.e. skewness and kurtosis are approximately zero), the number of sub-samples required to estimate the local variance at each sampling site at, say at a 10% ME and 95% statistical significance would be about 224 in all three examples (Appendix). In a more realistic scenario, however, the larger the local variance the greater the skewness and kurtosis are likely to be, and thus an even greater number of samples would be required. This re-emphasises the need to have a large survey's sampling density, so that local variances are not only small but also approximately more normal. As seen above, a less than ideal number of samples used in the estimation of the local variance tends to underestimate it, which can be highly misleading. The estimation of local variances requires that a huge number of sub-samples are not only sampled but also analysed individually (as opposed to local averages which can be analysed as a composite sample, Wolterbeek et al, 2010). A reasonable solution might be to analyse all sub-samples in a randomly selected fraction of the sampling sites, but this of course introduces a new source of uncertainty. This latter approach has become routine in atmospheric biomonitoring surveys, where sub-samples in about 20% of the sampling sites are analysed individually, whereas sub-samples in the remainder 80% of the sampling sites are analysed as composite samples (Jeran et al, 2003; Freitas et al, 1999; Ruhling, 1994). Obviously we have assumed throughout that all sampling sites have identical local variances,

when in reality this is seldom the case, however, so long as variances at each sampling site do not exceed the chosen maximum value, a margin of error for the survey's estimates can be confidently used.

The survey variance-to-local variance ratio (SV/LV) is a necessary and sound measure of the survey's quality (Wolterbeek et al, 1996) so long as both the survey's and the local variances are well estimated. As has been shown here and elsewhere (Wolterbeek & Verburg, 2002, 2004b; Wolterbeek et al, 1996), local variances estimated with less than the optimal number of sub-samples tend to be underestimated and thus reported SV/LV-ratios are probably inflated. In biomonitoring surveys at least, where only 5 sub-samples were usually used to estimate the local variance, this is probably the case (Jeran et al, 2003; Freitas et al, 1999; Ruhling, 1994). This understandable neglect derives from the fact that the survey's variance is the parameter of greatest direct interest and the fact that in order to estimate both the survey's and local variances one needs a multiplicative number of samples that need to be sampled and analysed individually (as opposed to composite sampling which can be used for measuring local averages). Since, even in the best of circumstances (local populations are normal) the total number of samples will be too large for the sampling and/or analytical capabilities of most research budgets and schedules, alternative or complementary strategies need to be developed in order to obtain accurate and precise estimates of local variances with less samples and/or less analyses. Composite sampling allied to large-sample analyses (Wolterbeek et al, 2010) are of no avail for estimating local variances, though they are very useful for estimating local averages. However, as seen above the advantages of this approach dim considering that the lower the local variances the lower the number (and size) of the sub-samples required to estimate the local average and thus the number/size/volume of composite samples will actually be relatively small. A more promising strategy to accurately determine local variances with a lower number of samples is the so-called "nearby sites" approximation based on kriging discussed in Wolterbeek & Verburg (2002, 2004b) and Wolterbeek et al (1996). And possibly the relationships found in a seminal elegant study which has shown that the "mean predicts the number of deviant individuals", at least in some data (Rose & Day, 1990).

The estimation of local populations also requires a concrete description of what is a sub-sample and what is their uncertainty (Aboal et al, 2006). Sub-sample uncertainty may be the analytical one or may be that stemming from sampling only a part of the sub-sample (e.g.: a fraction/mass of a particular lichen present among others in a particular tree). In this respect, the Ingamell's constant may be used (reviewed by Wolterbeek et al, 2010)

The discussion has focused on the more common case where the estimate of final interest is the population's variance. There are situations, where one may be more interested in higher moments (e.g.: surveillance of peaks or troughs, correlation analyses). As seen above the higher the moments the more they are affected by unsuitable sampling methods, sparse survey's sampling density and by sampling variability in general and thus the greater the sample size (sampling sites and sub-samples) they will require.

In what concerns sampling surveys that measure multiple parameters simultaneously one may calculate the MSS (n in the quotation) for each parameter but in the words of Cochran (1977, p81): "More commonly, there is a sufficient variation among the n's so that we are reluctant to choose the largest, either from budgetary consideration or because this will give an overall standard of precision substantially higher than originally contemplated. In this event, the desired standard of precision may be relaxed for certain of the items, in order to permit the use of a smaller value of n".

# 4 Geographical association of trace metal elements, measured by atmospheric biomonitoring, with circulatory diseases in the Portuguese population

*Based on article:*
*Sarmento SM., Verburg TG, FreitasMC & Wolterbeek HTh.*
*Submitted to Inhalation Toxicology,*
*January 2012.*

## 4.1 Abstract

Context and Objectives: This is the first complete epidemiological study to use atmospheric biomonitoring data as an indicator of human exposure to trace metal elements. The objectives are: 1) to estimate the chronic effect of trace metal elements on cardiovascular disease in the Portuguese population, and 2) to scrutinise significant relationships with regards to: estimation uncertainty, variable selection, and model selection uncertainty.

Materials/Methods: An aggregate ecological design, using 125 municipalities in Portugal, compared the concentration of 40 pollutants, with 11-years hospital admissions due to cardiovascular diseases. Single-pollutant linear models, F-change variable selection appropriate for confounding control, and the non-parametric bootstrap were used.

Results: Nearly all elements have an adverse effect. Arsenic, Nickel and Vanadium, as well as Magnesium, Potassium and Iron figure as conspicuous predictors of circulatory disease. Elasticities at the mean averaged 14% over all chemical elements. Standard errors were significantly underestimated in about half of the relationships, but in most cases this did not jeopardise statistical significance. Variable selection was extremely sensitive to sampling variability but was consistent within diagnostic-gender-age categories. The bootstrap did not appear to provide a satisfactory assessment of model selection uncertainty.

Conclusion: Atmospheric biomonitoring data is a promising tool for health impact assessment. The type of pollutants and the magnitude of their effect were consistent with previous epidemiological studies of metals. The bootstrap method is recommended for the assessment of precision and variable selection but not for the assessment of model selection uncertainty.

## 4.2 Introduction

The health effects associated with the chemical composition of air pollution, especially metals, was deemed a priority research topic by prominent health and environmental authorities (HEI, 2002; NRC, 1998). The emphasis on metals has been mostly motivated by: 1) toxicological studies that disclosed plausible biochemical pathways for cardiorespiratory toxicity (Ghio, 2004; Gavett et al, 2003; Knaapen et al, 2002; Lighty et al, 2000; Costa &

Dreher, 1999; Costa, 1998); 2) the long record of occupational epidemiology (e.g.: ATSDR, 2011). Epidemiological studies of environmentally-exposure to trace metal elements have been somewhat hampered by sparse or inexistent monitoring data, even those that are a recognised threat to human health (e.g.: As, Cd, Ni and Pb) (WHO, 2000; EIONET, 2011). The few extant studies have been mostly of a time-series design (e.g.: Dusseldorp et al, 1995; Thurston et al, 2005; Claiborn et al, 2002; Laden et al, 2000; Mar et al, 2006; Ito et al, 2006) and less than a handful of a cross-sectional design (Harrison et al, 2004; Lipfert et al, 2006). Cross-sectional studies require denser and wider monitoring networks than time-series studies. This is where organisms such as lichens and mosses, which feed mostly from atmospheric deposition, may be valuable. Atmospheric biomonitoring has a long history, being used not only for mapping but also to identify and locate emission sources of pollution (Ruhling, 1994; Garty et al, 2009; Buse et al, 2003; Harmens et al, 2004; see references in Wolterbeek et al, 2010). The main advantages of atmospheric biomonitoring relative to instrumental monitoring is the considerably lower costs and man-power required to: 1) perform high-density sampling over wide lengths of space and/or time; 2) monitor several pollutants simultaneously; and 3) ability to reflect cumulative exposures to air pollution and pre-instrumental pollution history. The disadvantages include: 1) a wide variety of factors can affect the accumulation of pollutants (e.g.: wind direction, weather, species, physiology, age and health); 2) biomonitoring reflects not only general atmospheric deposition but also local atmospheric sources such as soil re-suspension and non-atmospheric sources such as leachates from tree parts; 3) the time-interval over which pollutants are accumulated is unresolved, having been shown to range from two months to three years, depending on the trace metal element and environmental factors (Wolterbeek, 2002; Wolterbeek et al, 2010; Sloof & Wolterbeek, 1993a, 1993b; Reis et al, 2002; Godinho et al, 2004; Godinho, Verburg et al, 2009; Godinho, Wolterbeek et al, 2009; Godinho et al, 2011; Marques et al, 2004). Although modern atmospheric biomonitoring can, to some extent, account for or control some of the above sources of uncertainty (e.g.: by adjusting concentrations to measured environmental factors or by using biomonitor transplants), the lichen data used in this study was made prior to these developments.

The use of lichens and mosses in epidemiological research began with the influential Nature paper by Cislaghi & Nimis (1997), which reported a remarkable geographical correspondence between a lichen biodiversity index and the incidence of lung cancer in a northern region in Italy. However, this and subsequent studies have been mostly of an exploratory nature, because they used correlation measures and performed shy if any attempts at controlling confounding (Wappelhorst et al, 2000; Wolterbeek & Verburg, 2004a; Sarmento et al, 2008).

To the best of our knowledge, the present study is the first to use lichen monitoring data in combination with established epidemiological methods, including estimation of effect measures and confounding control. Despite these improvements, the results of this paper should be interpreted with caution, because atmospheric biomonitoring is an indirect indicator of atmospheric composition and the design is aggregate ecological. This study used a cross-

sectional design with 125 municipalities in Continental Portugal as the units of analysis and hospitalisations due to circulatory diseases as the health outcome.

This article has two aims. First, provide tentative estimates of the health effects of trace metal elements in atmospheric deposition. Second, assess uncertainties in estimation of effects, in particular: 1) robustness of estimates to sampling variability; 2) robustness of estimates to confounder selection; and 3) robustness of confounder selection to sampling variability and to variable's categories.

## 4.3 Methods

### 4.3.1 Hospital admissions database

The hospital admissions database was kindly provided by the Administração Central do Sistema de Saúde (ACSS, Portugal). It contained hospitalisation counts in public hospitals, summed over the years 1994-2004, for the 278 municipalities that form Continental Portugal (Figure 4.2). Hospital admissions were disaggregated by three diagnostic categories (acronym and ICD9-CM): circulatory disease (CIRC, 390-459), ischemic heart diseases (IHD, 410-414), and cerebrovascular diseases (CBV, 430-438); as well as by gender and three age-groups: 25-44, 45-64 and >64. Ischemic heart and cerebrovascular diseases were very rare in females 25-44 of age (Table 4.1) and thus were not considered in the analyses. In total, 16 diagnostic-gender-age categories were selected for analysis.

Table 4.1 Descriptive statistics (mean, standard deviation, minimum and maximum) of the 11-year prevalence per 1000 inhabitants, of hospital admissions in three diagnostics and six gender-age categories. Statistics were calculated over 227 sampling sites, corresponding to 125 municipalities.

| | | Females | | | | Males | | | |
|---|---|---|---|---|---|---|---|---|---|
| Age | Disease | Mean | SD | Min | Max | Mean | SD | Min | Max |
| 25-44 | CIRC | 29.96 | 12.78 | 6.30 | 76.96 | 22.24 | 6.37 | 7.76 | 39.47 |
| | IHD | 0.75* | 0.80 | 0.00 | 6.78 | 4.09 | 2.21 | 0.00 | 13.16 |
| | CBV | 2.51* | 1.22 | 0.00 | 6.88 | 2.86 | 1.16 | 0.00 | 8.01 |
| 45-64 | CIRC | 72.17 | 19.05 | 35.80 | 126.01 | 102.96 | 25.74 | 45.31 | 191.00 |
| | IHD | 10.18 | 4.77 | 0.00 | 26.44 | 35.93 | 13.21 | 9.84 | 88.94 |
| | CBV | 15.68 | 4.64 | 4.94 | 28.79 | 26.15 | 7.32 | 8.54 | 48.91 |
| >64 | CIRC | 246.76 | 70.32 | 83.77 | 461.54 | 335.10 | 84.49 | 157.38 | 600.06 |
| | IHD | 35.95 | 14.70 | 10.88 | 99.52 | 68.77 | 23.10 | 22.84 | 170.72 |
| | CBV | 108.22 | 29.28 | 36.44 | 191.77 | 133.82 | 34.82 | 48.05 | 248.49 |

* Hospital admissions excluded from analysis for being too rare. CIRC- circulatory diseases; IHD – ischemic heart diseases; CBV – cerebrovascular diseases.

The 16 hospital admission categories were standardised by the resident population and multiplied by 1000 (INE, 2011). The data provider was unable to exclude repeated admissions by the same patient caused by the same diagnostic, when the admissions took place in the same hospital (personal communication with Dr. Teresa Boto, ACSS, Portugal). For this reason, the standardised hospital admissions are probably best defined as 11-year prevalence per 1000 inhabitants (Table 4.1).

### 4.3.2 Trace metal elements database

Environmental exposure to atmospheric trace metal elements was assessed indirectly with lichen monitoring. The latter reflects the composition of atmospheric deposition and of non-atmospheric sources (e.g.: soil re-suspension, leachate from tree leaves). Existing studies suggest that the correlation between biological and instrumental monitoring tend to correlate moderately well in most cases, but this is issue requires more research (reviewed by Wolterbeek, 2002).

The concentration of 32 trace metal elements ($\mu g\ g^{-1}$ lichen) was obtained from a biomonitoring survey that sampled the lichen *Parmelia sulcata* on olive trees in the summer of 1993. Sampling was performed at 228 sampling sites throughout Continental Portugal (black squares, Figure 4.2). A more detailed account of the sampling and analytical procedures may be found in Reis (2001), Reis et al (1996) and Freitas et al (1997, 1999, 2000).

The trace element database was processed by Monte Carlo Target Transform Factor Analysis (MCTTFA), which identified eight emission sources (Kuik, Blaauw et al, 1993; Kuik, Sloof & Wolterbeek, 1993; Kuik & Wolterbeek, 1995).

Table 4.2 shows descriptive statistics for a selection of the 32 trace metal elements and eight emission factors found to be significant predictors of hospital admissions in single-pollutant models. Of the eight emission factors identified, three (F1, F2 and F5) were found to be significant predictors of hospital admissions. F1 appears to indicate a soil source since it contributes to a large fraction of the occurrence (approx. 30%) of a wide number of soil-related elements: Sc, Fe, Ti, Th and Sm, and it tends to concentrate in the mostly rural east. F2 is associated with a fuel combustion source, since it contributes greatly to the occurrence of Ni and V (approx. 50%) followed by I, Pb and Sb (approx. 30%) and its geographical distribution is consistent with urban and industrial locations. F5 appears to be a mixed factor, associated partly with a sea source and partly with an As source. It contributes substantially towards the occurrence of just three elements: Cl, Na and As (approx. 45%). Its geographical distribution is fairly homogeneous along the coast, consistent with a sea source, with some hotspots in the interior, possibly associated with As-rich soils or with the use of As-based pesticides in vineyards (Freitas et al, 1999, 2000) (Figure 4.1).

To ease readability, trace metal elements and emission factors are collectively referred to as pollutants in parts of the text.

Table 4.2 Descriptive statistics (mean, standard deviation, minimum and maximum) of the concentration (μg $g^{-1}$ lichen) of chemical elements and their associated emission factors (prefix F) determined in the lichen *Parmelia sulcata* in 227 sampling sites, corresponding to 125 municipalities in Continental Portugal (Figure 4.1). Only those chemical elements and emission factors that were found to be significant predictors of hospital admissions are shown.

| | Mean | SD | Min | Max | N(s)† | N(m)‡ | B/W§ |
|---|---|---|---|---|---|---|---|
| As | 1.72 | 0.93 | 0.71 | 4.85 | 213 | 115 | 2.31 |
| Cl | 1364 | 480 | 528 | 3200 | 227 | 125 | 0.94 |
| Cr | 5.26 | 2.28 | 1.96 | 13.40 | 219 | 118 | 1.16 |
| Cs | 0.60 | 0.30 | 0.22 | 1.72 | 213 | 115 | 1.72 |
| Eu | 0.18 | 0.08 | 0.07 | 0.48 | 217 | 118 | 1.03 |
| Fe | 2125 | 989 | 705 | 5320 | 216 | 118 | 1.46 |
| Hf | 0.41 | 0.20 | 0.14 | 1.12 | 210 | 113 | 1.37 |
| I | 6.78 | 3.18 | 2.24 | 17.60 | 222 | 122 | 1.57 |
| K | 5462 | 1561 | 2280 | 10 900 | 227 | 125 | 1.61 |
| La | 2.98 | 1.46 | 1.01 | 7.80 | 210 | 114 | 1.10 |
| Mg | 1986 | 741 | 772 | 4690 | 224 | 124 | 2.12 |
| Mn | 51.05 | 17.91 | 18.70 | 115 | 226 | 124 | 1.69 |
| Ni | 3.76 | 2.06 | 1.33 | 10.60 | 199 | 112 | 1.76 |
| Sb | 0.29 | 0.16 | 0.11 | 0.84 | 205 | 112 | 2.38 |
| Sc | 0.40 | 0.14 | 0.16 | 1.02 | 224 | 124 | 1.27 |
| Sm | 0.44 | 0.20 | 0.15 | 1.16 | 218 | 117 | 1.06 |
| Th | 0.88 | 0.46 | 0.32 | 2.48 | 205 | 111 | 1.81 |
| Ti | 330 | 152 | 111 | 808 | 220 | 121 | 1.16 |
| V | 13.99 | 7.87 | 5.35 | 40.60 | 206 | 116 | 1.57 |
| F1 | 62.28 | 64.84 | 0.00 | 362 | 227 | 125 | 2.20 |
| F2 | 10.07 | 8.21 | 0.00 | 47.23 | 227 | 125 | 1.41 |
| F5 | 38.91 | 18.82 | 8.65 | 109.50 | 227 | 125 | 1.50 |

† Number of non-missing sampling sites. ‡ Number of municipalities. § Between-area to within-area variance ratio.

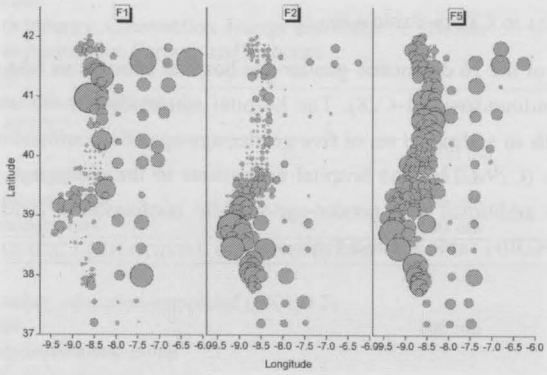



Figure 4.1Geographical distribution of the values of the emission factors (F1, F2, F5) found to be significant predictors of hospitalisations. The bubble's area is proportional to the factors' values and their location corresponds to the coordinates of the 228 lichen sampling sites (black squares in Figure 4.2).

### 4.3.3 Confounders database

The confounders database was assembled from two sources: the website (INE, 2011) and a CD-ROM publication (INE, 2006) of the Portuguese National Statistical Institute (INE, Portugal).

The selection of potential confounding variables was made by benchmarking state-of-the-art multi-level prospective cohort studies of air pollution, assisted by substantive reasoning regarding the variables' causal proximity, data quality (e.g.: extreme values, variance), susceptibility to measurement error (e.g.: information bias) and construct redundancy (e.g.: Krewski et al, 2000; Willis et al, 2003; Jerrett et al, 2003; Lipfert & Morris, 2002; Lipfert et al, 2000; Lipfert et al, 2003; Sarmento et al, 2008). In an attempt to minimise confounding by group and effect modification by group, confounders were selected on the basis that they were risk factors for disease, regardless of their correlation with the elemental concentrations between groups (Morgenstern, 2008; Willis et al, 2003; Greenland & Morgenstern, 1989). Lifestyle variables and physiological parameters were only available for regions larger than municipalities and thus could not be used (INE & INSA, 2009).

Twenty-eight variables were selected that covered ten broad types of health determinants: air pollution sources, exposure assessment, places, demography, socioeconomics, social capital, employment, education, living conditions and health status/services (C1 to C28 in Table 4.3 and Table 4.4). These 28 variables refer to characteristics of the municipalities or of the total population within each municipality, regardless of gender and age. Since the hospital admissions data refer to specific gender-age categories, the lack of equivalent standardisation of some confounders could cause a bias known as mutual standardisation (Rosenbaum & Rubin, 1984). From the census of 2001, it was possible to obtain some confounder data disaggregated by gender-age categories, namely: marital status, education and employment (C29 to C33 in Table 4.3) (INE, 2011). Employment data was only available for working age-groups and was represented by three variables: unemployment, and employment in industry and agriculture (C31 to C33 in Table 4.3).

In summary, each of the 16 diagnostic-gender-age hospital admissions was modelled with the same set of 28 confounders (C1-C28). The hospital admissions in the working-age groups were modelled with an additional set of five gender-age-specific confounders, yielding a total of 33 confounders (C29-C33). The hospital admissions in the oldest age-group (>64) were modelled with an additional two gender-age-specific confounders, yielding a total of 30 confounders (C29-C30) (Table 4.3 and Table 4.4).

Table 4.3 Confounders included in the Full model, divided into 10 types of health determinants (capital letters) and into confounders describing the total-population and confounders describing specific gender-age groups of the population. Inh=inhabitant.

| Code | Category and name | Units | Time |
|---|---|---|---|
| TOTAL POPULATION ECOLOGICAL CONFOUNDERS | | | |
| | AIR POLLUTION SOURCES | | |
| C1 | Gasoline sales | Tonnes/km$^{-2}$ | 1996-04 |
| C2 | Industrial Electricity Consumption | % total electricity consumption | 1998-03 |
| | SOCIO-ECONOMICS | | |
| C3 | Credit | 1000 €/inh | 1995-04 |
| C4 | Tax on motor vehicles | 1000€/inh | 1995-03 |
| C5 | Purchasing power | Inh | 1993-04 |
| | SOCIAL CAPITAL | | |
| C6 | Survival pension | 100 inh | 1994-04 |
| C7 | Crimes against people | 1000 inh | 1998-00 |
| C8 | Divorced | 100 inh | 2001 |
| | PLACES | | |
| C9 | Population density | inh/km$^{-2}$ | 1997 |
| C10 | Places<2000 inhabitants | 100 inh | 2001 |

Table 4.3 continued

| Code | Category and name | Units | Time |
|---|---|---|---|
| | LIVING CONDITIONS | | |
| C11 | Domestic electricity consumption | kW$^{-1}$ per hour per 1000 inh | 1998-00 |
| C12 | Water consumption | M$^{-3}$ per inhabitant | 2001-03 |
| C13 | Urban waste | Kg$^{-1}$ per inhabitant | 2002-04 |
| C14 | Central Heating | % housing | 2001 |
| C15 | Classical Families with one person | % classical families | 2001 |
| | HEALTH STATUS | | |
| C16 | Infant mortality | 1000 live births | 1998-02 |
| C17 | Appointments in Health Centres | Inh | 1994-03 |
| | EXPOSURE ASSESSMENT | | |
| C18 | Immigrants from other municipalities | 100 inh | |
| C19 | Emigrants to other municipalities | 100 inh | 2001 |
| C20 | Commuting – mean duration | minutes | |
| | EMPLOYMENT | | |
| C21 | Unemployment | | |
| C22 | Employed in Industry, Construction, Energy and Water | 100 inh | 2001 |
| C23 | Employed in Agriculture, Forestry and Fisheries | | |
| | EDUCATION | | |
| C24 | Compulsory Education (9years) | 100 inh | 1991 |
| | DEMOGRAPHY | | |
| C25 | Birth rate | 1000 inh | 1992-94 |
| C26 | Population change | 100 inh | 1991-01 |
| C27 | Child dependency ratio | 100 inh | 1991 |
| C28 | Aged dependency ratio | 100 inh | 1991 |
| GENDER-AGE-SPECIFIC ECOLOGICAL CONFOUNDERS | | | |
| C29 | Divorced | | |
| C30 | Lower secondary education-completed (ISCED 2) | | |
| C31 | Unemployed | 100 inh | 2001 |
| C32 | Industrial socioeconomic group | | |
| C33 | Agricultural socioeconomic group | | |

Table 4.4 Descriptive statistics (mean, standard deviation, minimum and maximum) of the confounders (see Table 4.3). Statistics were calculated over 227 sampling sites, corresponding to 125 municipalities. Statistics for the confounders C29 to C33 (Table 4.3) are not shown because they have different values depending on the gender-age group.

| | Mean | SD | Min | Max | | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|---|---|---|
| C1 | 33.97 | 74.64 | 0.10 | 906.80 | C15 | 17.31 | 4.50 | 6.81 | 27.90 |
| C2 | 37.27 | 20.40 | 3.77 | 92.82 | C16 | 4.81 | 2.50 | 0.00 | 12.60 |
| C3 | 5.67 | 2.71 | 1.17 | 16.45 | C17 | 2.85 | 0.57 | 1.78 | 5.21 |
| C4 | 0.58 | 0.14 | 0.26 | 1.26 | C18 | 5.64 | 3.13 | 2.23 | 20.77 |
| C5 | 71.10 | 21.37 | 35.89 | 161.00 | C19 | 4.83 | 1.67 | 2.80 | 12.39 |
| C6 | 6.21 | 1.53 | 3.26 | 11.39 | C20 | 18.61 | 4.82 | 11.98 | 38.24 |
| C7 | 4.41 | 1.98 | 1.23 | 13.53 | C21 | 6.57 | 2.31 | 2.50 | 16.50 |
| C8 | 1.53 | 0.57 | 0.50 | 4.00 | C22 | 37.95 | 11.63 | 14.58 | 65.66 |
| C9 | 189.81 | 334.21 | 7.66 | 3308.02 | C23 | 8.37 | 6.52 | 0.43 | 35.09 |
| C10 | 60.78 | 24.45 | 5.68 | 98.73 | C24 | 16.43 | 6.65 | 6.05 | 50.29 |
| C11 | 857.20 | 131.70 | 573.46 | 1337.87 | C25 | 10.22 | 2.52 | 5.23 | 19.17 |
| C12 | 63.66 | 33.10 | 8.33 | 190.00 | C26 | 4.78 | 9.94 | -17.95 | 39.39 |
| C13 | 406.98 | 117.05 | 168.49 | 811.37 | C27 | 29.86 | 4.01 | 22.40 | 46.30 |
| C14 | 5.78 | 4.64 | 0.10 | 19.30 | C28 | 24.43 | 8.09 | 9.70 | 66.80 |

### 4.3.4 Study area and unit of analysis

The epidemiological design was aggregate ecological with a subset of municipalities in Continental Portugal as the unit of analysis and study area, respectively. Municipalities were the lowest geographical unit for which health data were freely and readily available (personal communication with Dr. Teresa Boto, ACSS, Portugal and INE, Portugal), and were also the lowest geographical unit for which many confounders were available (INE, 2011, 2006). The hospitalisation period 1994-2004 was chosen because it starts the year after the lichen survey and it ends before major changes in the way hospitalisation are recorded were made (personal communication with Dr. Teresa Boto, ACSS, Portugal and INE, Portugal).

While the hospitalisation and confounders database covered all 278 municipalities that make up Continental Portugal, the trace metal elements database, consisted of 228 sampling sites that covered only 126 municipalities in Continental Portugal (Figure 4.2). One municipality (Loulé, Algarve) had to be excluded because it showed extreme values for several confounders, likely owing to its intense touristic activity. The intersection of the three databases thus yielded 125 unique municipalities on which analysis could be carried out (grey polygons, Figure 4.2).

Statistical interpolation of the pollutant's 227 point concentrations into surface concentrations matching the 125 municipalities was not performed because for most pollutants the spatial dependence was either weak or undetectable (not shown; Cressie, 1993; Wakefield, 2004; Wakefield & Shaddick, 2006; Wolterbeek & Verburg, 2004b). Instead, all point concentrations were used in regression analyses (see below).

The hospital admissions and confounders database were adjusted so that municipalities with more than one lichen sampling site, had repeated values. For example, if a municipality contained three sampling sites, the municipality was represented by three observations, all with the same value for the hospital admissions and confounders, but different values for the chemical elements corresponding to the concentrations recorded at each of the three sampling sites. This procedure amounts to a weighed regression that accounts for the observed within-area variability in pollutant concentrations. Thus the effective number of observations (i.e. N) used in regression analysis was 227, corresponding to 125 unique municipalities.

Table 4.5 shows some of the characteristics of the study area. The between-area to within-area variance ratio (B/W) (also known as signal-to-noise ratio) (Salway, 2003; Wolterbeek & Verburg, 2002) of the pollutants was calculated by averaging the within-area variance over the 51 municipalities with more than one sampling site and by calculating the between-area variance over the mean values of the 125 municipalities.

Figure 4.2 Geo-referenced map depicting the 278 municipalities that form Continental Portugal (grey and white polygons). Black squares represent the 227 sampling sites where chemical elements were measured. Grey polygons represent the study area, i.e. 125 municipalities with at least one sampling site (black squares).

Table 4.5 Descriptive statistics (median, minimum and maximum) of the study area (Figure 4.1). Also shown is the total number of hospital admissions over the study period the study period (1994-2004).

| | Median | Min | Max | % of Portugal |
|---|---|---|---|---|
| Area ($km^{-2}$) | 245 | 21 | 1722 | 51% |
| Inhabitants | 23 389 | 3393 | 363 749 | 56% |
| CIRC | 1886 | 324 | 21 971 | 52% |
| IHD | 331 | 48 | 5975 | 50% |
| CBV | 682 | 118 | 6736 | 52% |
| Sampling sites | 1.8 | 1 | 10 | NA |
| B/W† | 1.55 | 0.94 | 2.38 | NA |

† Between-area to within-area variance ratio.

### 4.3.5 Pre-selection of relationships

The databases, consisting of 16 diagnostic-gender-age hospital admission categories and 40 pollutants resulted in no less than 640 possible single-pollutant relationships.

Selection of the most significant relationships was performed using two criteria. The first required that the F-value of the model and the t-value of the pollutant were both highly significant at $p<.01$ in the Full model (i.e. model that contains a pollutant and all 30-33 confounders). The second required that the t-value of the element/factor was moderately significant at $p<.05$ in the Simple model (i.e. model that contains a pollutant and no confounders). The first criterion is typical of environmental epidemiological studies (e.g.: Lipfert et al, 2000), whereas the second is meant to avoid cases of complete negative suppression, which challenge causal interpretation and may show inflated statistical significance (Friedman & Wall, 2005; Tzelgov & Henik, 1991). The criteria selected 50 relationships on which all analyses will be performed.

### 4.3.6 Non-parametric bootstrap

The non-parametric bootstrap was used to assess the robustness of estimation and of variable selection to sampling variability, and to assess the robustness of estimation to multiple testing (model selection uncertainty). This technique generates (bootstrap) samples by sampling-with-replacement from the vector of observations over all variables required for regression (hospitalisations, pollutants and confounders) (Efron & Tibshirani, 1993; Stine, 1989). Note that the bootstrap relies on a single but un-checkable assumption: that the bootstrap sampling distribution is able to reproduce the true sampling distribution (Efron & Tibshirani, 1993; Stine, 1989; Chernick, 2008).

One thousand (500 in section "Model selection uncertainty") bootstrap samples were generated. The number of observations in bootstrap samples equalled that in the original sample (N=227 if no observations were missing, Table 4.2). The mean and standard deviation of the bootstrap sampling distributions correspond to the bootstrap slope (B*) and the bootstrap standard error (SE*), respectively (Efron & Tibshirani, 1993; Stine, 1989).

### 4.3.7 Linear regression and estimation of effects

Ordinary Least Squares (OLS) linear regression was performed on both observed sample and bootstrap samples. Linear regression was used for three reasons. First, all variables are continuous and the study design is aggregate ecological; in such cases, linear regression has been recommended (Rothman, 2002; Greenland, 1992; Greenland & Robins, 1994; Salway, 2003; Glynn et al, 2008). Second, the study area is Portugal, a fairly small and un-industrialised country which benefits from favourable dominant Atlantic winds. Thus it is reasonable to assume that exposure is low and has a narrow range, relative to the full exposure range of the true dose-response curve. In such cases, a linear approximation is reasonable (Rothman, 2002; Wakefield, 2003; Salway & Wakefield, 2004). Third, non-linear

models assume that the exposure effect interacts with the confounder, which seems unreasonable, especially in an ecological study (Rothman, 2002).

Under strong causal assumptions (Rothman, 2002), the linear slope of the pollutant ($\beta_x$) is assumed to be equivalent to the effect estimate Risk Difference (RD). The RD was converted into relative measures of effect: Risk Ratio (RR) and Elasticity at the mean (E). RD and RR were expressed in relation to an "achievable change" (a) in pollutants, i.e. the mean minus the minimum concentration of the pollutants across the 125 municipalities (Lipfert et al, 2006). Thus:

$$\text{aRD} = \beta_x \left[\text{mean}(X) - \min(X)\right],$$

$$\text{aRR} = \frac{\alpha + \beta_x.\text{mean}(X) + \sum_{i=1}^{k} \beta_{z_k} \text{mean}(Z_k)}{\alpha + \beta_x.\min(X) + \sum_{i=1}^{k} \beta_{z_k} \text{mean}(Z_k)} \text{ and}$$

$$\text{E} = \beta_x \frac{\text{mean}(X)}{\text{mean}(Y)}$$

(Lipfert et al, 2006; Cohen et al, 2003; Baxter et al, 1997; Cameron & Trivedi, 1989; Greenland & Morgenstern, 1989). E was calculated because it is more readily comparable with the results of existing epidemiological studies (Lipfert, 1993; Baxter et al, 1997; Lipfert et al, 2006).

Estimates are presented for the F model only because it includes all confounders selected a priori on the basis of substantive reasoning (Chen et al, 1999; Jorgensen et al, 2007; Fewell et al, 2007; Robins & Morgenstern, 1987).

### 4.3.8 Estimation uncertainty

Assessment of the robustness of estimation to sampling variability was performed by comparing the pollutant's slope and standard error estimated in the observed sample (called naïve estimates and dented by B and SE) with those estimated from the bootstrap sampling distribution (called bootstrap estimates and denoted by B* and SE*). Comparisons were expressed in terms of %bias: the difference between the naïve and bootstrap estimates, relative to the naïve estimate. By convention, the bias was judged significant when it reached >|10%|. These comparisons were performed for the estimates obtained from each of the four model specifications (see next section).

### 4.3.9 Model reduction

Four model specifications were estimated with OLS linear regression. All models included a single pollutant, which was always force-entered in the model, and 33 confounders (30 for hospital admissions in the >64 age-group), which were subject to variable selection procedures. The four model specifications are listed below.

1. Full model (F) – all 30-33 confounders selected a priori were force-entered in the model.
2. Backward model (B) – confounders were selected backwards, starting from the F model, using the F-change criterion at $p<.20$.
3. Mean Backward model (MB) – confounders were selected by model averaging (Chatfield, 1995), which involved two stages. First, for each relationship 500 bootstrap samples were generated and the B model was applied to each bootstrap sample. Second, the mean (k) number of confounders selected across the 500 bootstrap samples was calculated and the most prevalent k confounders identified. The MB model for each relationship was then specified by these k most recurrent confounders.
4. Simple model (S) – no confounders were entered in the model.

Confounders in B and MB models were selected backwards with F-change at $p<.20$. This criterion was chosen because it has been shown to be an adequate alternative to the Change in Estimate (CE) criterion with a cut-off of 10%. The potential advantage of F-change over CE is that it prevents both important confounders and important predictors from being excluded (Mickey & Greenland, 1989; Maldonaldo & Greenland, 1993; Greenland, 1989; Jorgensen et al, 2007).

Over the 50 relationships, the MB and B model contained on average 20 (min-max: 17-23) and 18 (12-22) confounders, respectively.

The slope (B), standard error (SE) and multiple correlation coefficient ($R^2$) obtained from the B, MB and S models were compared with those obtained from the F model. The comparison was expressed in % bias: difference between the F model's estimate and the reduced model's estimate, relative to the F model's estimate. By convention, a significant bias was judged when it reached $>|10\%|$. A difference in slope of $>|10\%|$ between alternative models is commonly used as an indicator of residual confounding (e.g.: Jorgensen et al, 2007; Fewell et al, 2007; Robins & Morgenstern, 1987; Rothman, 2002).

B and MB models are collectively referred to as reduced models in the text, to contrast them with the F model.

### 4.3.10 Model selection uncertainty

Model selection uncertainty was assessed by a method suggested by Chatfield (1995), which is called here the V model, and is based on the non-parametric bootstrap. For each relationship, 500 bootstrap samples were generated. Variable selection was performed on each bootstrap sample using the B model (see above). The resulting sampling distribution's mean and standard deviation were used to determine the bootstrap slope (B*) and standard error (SE*). Note that the V model's estimates are different from the bootstrap estimates obtained from the other models (F, B and MB) because with the former, variable selection is performed on each bootstrap sample, whereas with the latter, variable selection is performed on the observed sample and then the selected model is fixed and fitted to each bootstrap sample.

The bootstrap slope (B*) and bootstrap standard error (SE*) obtained from the V model were compared with those obtained from the F, B and MB models. The comparisons were expressed in %bias: the difference between the V model's estimate and that of the other models, relative to the V model's estimate. By convention, a significant bias was judged when it reached >|10%|.

### 4.3.11 Robustness of confounder selection

Robustness of variable selection was evaluated relative to two aspects: sampling variability and within hospitalisation categories. Both evaluations used the confounder composition of each of the 500 models selected by the V model.

The robustness of variable selection to sampling variability was assessed by noting the frequency with which identical confounder combinations occurred across the 500 bootstrap samples. For instance, for a given relationship the V model may result in the selection of model A (e.g.: C1+C2+C5+C11+C29) in 50 bootstrap samples, model B (e.g.: C5+C11+C15) in 100 bootstrap samples, and model C (e.g.: C1+C2+C5+C11+C25+C27+C32) in 350 bootstrap samples. Then, the maximum frequency) was found (in the example 350).

The consistency of variable selection within and between hospitalisation categories (i.e. diagnostic-gender-age groups) was assessed by determining the inclusion frequency, i.e. number of times a confounder is included over the 500 models obtained from the V model (min: 0; max: 500) (Heymans et al, 2007). The inclusion frequency over all confounders was then correlated between relationships that belonged to the same hospitalisation category, but with different pollutants (i.e. within hospitalisation categories) and between relationships that did not belong the same hospitalisation category (i.e. between hospitalisation categories).

### 4.3.12 Software

SPSS 17.0 syntax was used to perform the non-parametric bootstrap and OLS linear regression. ArcGIS Explorer Desktop was used to plot the geo-referenced map in Figure 4.1.

## 4.4 Results

### 4.4.1 Estimation of pollutant effects

Table 4.6 presents the pollutant's effect as estimated with the F model, for each of the 50 relationships.

Interpretation of linear regression estimates as epidemiological measures of effect requires strong assumptions (Rothman, 2002), the most important of which are: 1) lichens reflect human inhalation exposure to trace metal element; 2) relationship between predictors and hospitalisations are causal; and 3) no biases are present, especially measurement error and residual confounding (Rothman, 2002; Morgenstern, 2008; Salway & Wakefield, 2004). This study cannot ensure compliance to these assumptions, thus Table 4.6 should be interpreted with great caution.

Table 4.6 Effect estimates of the pollutant in the 50 selected relationships, estimated by the Full model. Estimates are ordered by the Risk Ratio for an achievable change in pollutant (aRR).

| Relationships | $R^2$ | T¶ | A† | aRD‡ | aRR‡ | E§ | SE(E) § | SE*(E) § |
|---|---|---|---|---|---|---|---|---|
| F65+-IHD-As | 0.64 | 0.64 | 1.011 | 4.784 | 1.193 | 0.226 | 0.042 | 0.064 |
| F65+-IHD-F2 | 0.61 | 0.51 | 10.09 | 3.608 | 1.175 | 0.100 | 0.031 | 0.034 |
| M25-44-IHD-Mg | 0.43 | 0.63 | 1215 | 0.817 | 1.171 | 0.326 | 0.100 | 0.106 |
| F45-64-CBV-K | 0.49 | 0.68 | 3183 | 2.256 | 1.168 | 0.247 | 0.065 | 0.066 |
| F45-64-IHD-As | 0.57 | 0.63 | 1.011 | 1.341 | 1.149 | 0.224 | 0.052 | 0.064 |
| M25-44-IHD-Ni | 0.47 | 0.58 | 2.434 | 0.743 | 1.147 | 0.281 | 0.073 | 0.072 |
| F45-64-CBV-Se | 0.50 | 0.78 | 0.247 | 1.958 | 1.143 | 0.205 | 0.050 | 0.055 |
| F45-64-CBV-Mg | 0.50 | 0.69 | 1215 | 1.904 | 1.138 | 0.199 | 0.049 | 0.047 |
| M25-44-IHD-V | 0.46 | 0.54 | 8.642 | 0.603 | 1.132 | 0.238 | 0.072 | 0.074 |
| F45-64-CBV-Th | 0.56 | 0.65 | 0.557 | 1.786 | 1.131 | 0.180 | 0.036 | 0.045 |
| F65+-CIRC-As | 0.55 | 0.64 | 1.011 | 23.323 | 1.124 | 0.160 | 0.033 | 0.043 |
| M25-44-CBV-Sb | 0.40 | 0.48 | 0.176 | 0.336 | 1.114 | 0.191 | 0.062 | 0.078 |
| F45-64-CBV-Cs | 0.56 | 0.69 | 0.378 | 1.591 | 1.114 | 0.160 | 0.035 | 0.036 |
| F65+-IHD-Ni | 0.65 | 0.57 | 2.434 | 3.112 | 1.114 | 0.134 | 0.044 | 0.042 |
| M25-44-IHD-Fe* | 0.44 | 0.65 | 1421 | 0.702 | 1.112 | 0.257 | 0.081 | 0.099 |
| M25-44-IHD-Cr | 0.42 | 0.71 | 3.305 | 0.613 | 1.105 | 0.239 | 0.082 | 0.089 |
| F45-64-CBV-Mn* | 0.47 | 0.63 | 32 | 1.476 | 1.104 | 0.149 | 0.056 | 0.058 |
| M45-64-CBV-I | 0.44 | 0.46 | 4.543 | 2.172 | 1.103 | 0.124 | 0.047 | 0.048 |
| M25-44-IHD-F2 | 0.43 | 0.50 | 10 | 0.653 | 1.099 | 0.159 | 0.051 | 0.051 |
| M25-44-IHD-As | 0.43 | 0.62 | 1.011 | 0.514 | 1.094 | 0.213 | 0.071 | 0.077 |
| F45-64-CIRC-As | 0.55 | 0.63 | 1.011 | 6.181 | 1.094 | 0.145 | 0.031 | 0.034 |
| F45-64-CBV-Sm | 0.51 | 0.72 | 0.292 | 1.321 | 1.093 | 0.128 | 0.039 | 0.045 |
| F45-64-CBV-V | 0.49 | 0.51 | 8.642 | 1.266 | 1.088 | 0.131 | 0.041 | 0.047 |
| F45-64-CBV-Eu | 0.52 | 0.74 | 0.118 | 1.209 | 1.084 | 0.120 | 0.039 | 0.045 |
| F45-64-CBV-Fe* | 0.50 | 0.64 | 1421 | 1.208 | 1.084 | 0.115 | 0.042 | 0.050 |
| M45-64-CBV-F2* | 0.44 | 0.47 | 10 | 1.965 | 1.080 | 0.075 | 0.027 | 0.031 |
| F65+-CBV-Ni | 0.39 | 0.57 | 2.434 | 7.606 | 1.080 | 0.109 | 0.038 | 0.038 |
| F45-64-CBV-Ti* | 0.49 | 0.64 | 219 | 1.152 | 1.080 | 0.111 | 0.042 | 0.051 |
| F45-64-CBV-Sb* | 0.49 | 0.46 | 0.176 | 1.138 | 1.079 | 0.118 | 0.042 | 0.056 |
| F45-64-CBV-La | 0.52 | 0.76 | 1.969 | 1.135 | 1.079 | 0.110 | 0.037 | 0.039 |
| F65+-CBV-Cs | 0.44 | 0.69 | 0.378 | 8.508 | 1.078 | 0.124 | 0.036 | 0.038 |
| M45-64-CBV-Ni* | 0.41 | 0.55 | 2.434 | 1.903 | 1.077 | 0.112 | 0.039 | 0.046 |
| M45-64-CIRC-As | 0.53 | 0.63 | 1.011 | 8.448 | 1.077 | 0.139 | 0.030 | 0.036 |
| F65+-CBV-V | 0.41 | 0.55 | 8.642 | 8.508 | 1.077 | 0.127 | 0.038 | 0.043 |
| F45-64-CBV-F1 | 0.49 | 0.59 | 62 | 1.093 | 1.075 | 0.070 | 0.019 | 0.025 |
| F45-64-CIRC-F5* | 0.49 | 0.69 | 30 | 4.984 | 1.074 | 0.089 | 0.034 | 0.036 |
| F45-64-CBV-As | 0.50 | 0.63 | 1.011 | 1.078 | 1.074 | 0.117 | 0.037 | 0.044 |
| F65+-IHD-V* | 0.63 | 0.55 | 8.642 | 2.741 | 1.073 | 0.123 | 0.044 | 0.049 |
| M45-64-IHD-As | 0.55 | 0.63 | 1.011 | 3.618 | 1.072 | 0.171 | 0.043 | 0.053 |
| F45-64-CBV-Hf* | 0.53 | 0.76 | 0.269 | 1.028 | 1.071 | 0.101 | 0.035 | 0.041 |
| M65+-CIRC-As | 0.53 | 0.64 | 1.011 | 27.353 | 1.070 | 0.139 | 0.030 | 0.036 |
| F65+-CBV-As | 0.42 | 0.64 | 1.011 | 7.563 | 1.069 | 0.119 | 0.036 | 0.040 |

Table 4.6 Continued

| Relationships | $R^2$ | T¶ | A† | aRD‡ | aRR‡ | E§ | SE(E) § | SE*(E) § |
|---|---|---|---|---|---|---|---|---|
| M45-64-CBV-Cs* | 0.48 | 0.68 | 0.378 | 1.617 | 1.064 | 0.098 | 0.037 | 0.042 |
| M65+-IHD-As | 0.68 | 0.64 | 1.011 | 5.535 | 1.063 | 0.137 | 0.032 | 0.041 |
| M25-44-CIRC-As | 0.54 | 0.62 | 1.011 | 1.847 | 1.061 | 0.141 | 0.034 | 0.039 |
| M65+-CBV-Cs | 0.43 | 0.69 | 0.378 | 9.090 | 1.060 | 0.107 | 0.035 | 0.036 |
| M65+-CBV-Ni* | 0.38 | 0.59 | 2.434 | 8.723 | 1.056 | 0.101 | 0.037 | 0.040 |
| M65+-CBV-V* | 0.39 | 0.56 | 8.642 | 8.669 | 1.055 | 0.105 | 0.037 | 0.043 |
| M65+-CBV-As | 0.40 | 0.64 | 1.011 | 8.527 | 1.055 | 0.108 | 0.035 | 0.037 |
| M45-64-CBV-Cl | 0.44 | 0.81 | 837 | -2.374 | 0.919 | -0.148 | 0.048 | 0.047 |

* Relationships where element/factor ceases to be statistically significant at $p<.01$ when SE*(E) are used in place of SE(E). ¶ Tolerance. † Achievable change in element/factor. ‡ Risk different and risk ratio per achievable change. § Elasticity at the mean and its naïve (SE(E)) and bootstrap (SE*(E)) standard errors.

Over the 50 relationships, pollutants are always positively associated with hospital admissions, except for Cl and cerebrovascular disease in males 45-64 years old. As is the most prevalent pollutant, being significantly associated with 13 hospitalisation categories. Ni, V and F2 are very inter-correlated and appear to signal an urban-rural gradient, since they show high collinearity with the confounders (Tolerance, Figure 4.6). Strong effects were also observed with seemingly innocuous elements such as Mg, K and Fe, which are usually associated with lichen physiology.

The aRD is a measure of impact, as such it tends to be larger for hospitalisation categories that are common. On the average, 4 hospitalisations (min-max: -2-27) are associated with an achievable change in pollutant concentrations.

The aRR averages 1.09 (min-max: 0.92-1.19) suggesting a 9% excess risk of hospitalisations per achievable change in pollutant concentrations.

The E correlates well with aRR, and suggests an average 15% increase in hospitalisations (min-max: -15-33%) per 1% increase in pollutant concentrations.

There is no apparent pattern whereby particular types of pollutants (e.g.: urban/industrial origin such as Ni, V and F2) have a stronger effect (aRR and E) than other types of pollutants (e.g.: natural origin such as Mg, Fe and K).

### 4.4.2 Estimation uncertainty

The assessment of the robustness of estimates to sampling variability is important because estimates based on a single sample may be biased by violations of OLS assumptions (e.g.: outliers, heteroscedascity, non-normal and/or non-independent residuals) (Efron & Tibshirani, 1993; Stine, 1989; Chernick, 2008 Chapter 4). This assessment is made by generating sampling variability through the bootstrap.

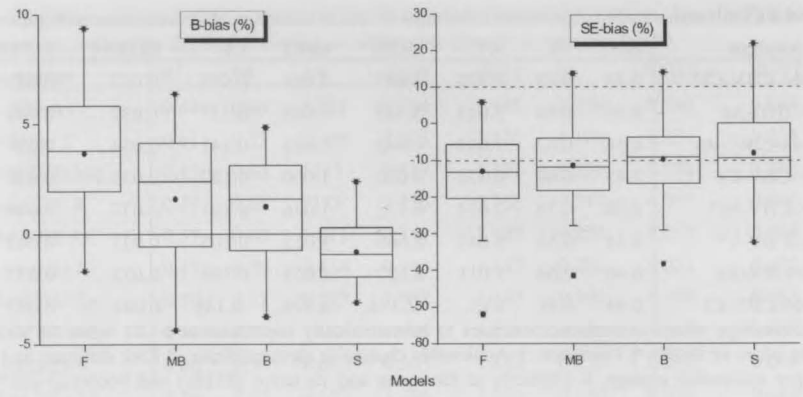



Figure 4.3 Boxplots (minimum, 25%, 50%, 75% and maximum), calculated over the 50 relationships, of the %bias between naïve and bootstrap estimates of the slope (B-bias) and the standard error (SE-bias). Results are shown for the four models (x-axis).

Figure 4.3 shows the estimation uncertainty for each of the four models (F, B, MB and S), averaged over the 50 relationships. The SE-bias, and to a lesser extent, the B-bias are fairly similar across the three models. Thus generalising across the four models: B-bias averages +1% (min-max: -3+6%) with all of the relationships having a B-bias<|10%|, whereas the SE-bias averages -11% (-41+14%) with 47% of the relationships having an SE-bias<|10%|.

The results indicate that sampling variability does not affect B estimation to any great extent. This was expected since the mean of the bootstrap sampling distribution should equal the sample's parameter, unless the sample size is too small (Efron & Tibshirani, 1993; Stine, 1989; Chernick, 2008). There is a clear tendency for B* to become larger than B as the model becomes larger, probably owing to the decreasing numbers of degree of freedom (Chatfield, 1995).

With regards to the SE, sampling variability causes SEs to increase in all relationships; however the increase is only substantial for about half of the relationships. In these cases, the SE* should be reported instead of SE. Using SE* instead of SE, leads to the pollutant's effect in 14 relationships losing statistical significance at $p<.01$ but not at $p<.05$ (asterisk in Table 4.6). Table 4.6 shows both SE and SE* for the elasticity at the mean estimated by the Full model. There is no obvious pattern whereby particular pollutants and/or hospitalisations are more greatly affected by sampling variability. The fact that SE of all four models are similarly affected by sampling variability suggests the presence of a problem in the data that none of the (nested) models is able to accommodate. The most straightforward problem is the presence of outliers (not shown) (Efron & Tibshirani, 1993; Stine, 1989; Chernick, 2008).

### 4.4.3 Model reduction

A large body of evidence has shown that the best model for causal inference is the one selected a priori on the basis of substantive reasoning (e.g.: Chen et al, 1999; Jorgensen et al, 2007; Fewell et al, 2007; Robins & Morgenstern, 1987). In the present work this model is

represented by the F model. However, such models tend to present large SE (Goldberger, 1991 p248; Jorgensen et al, 2007; Breiman, 1992; Friedman & Wall, 2005), and are susceptible to fitting error/noise and to finite sample-bias (Greenland, 1989; Salway, 2003 Chapter 3). Thus it may be worthwhile to consider model reductions. The aim is to examine whether the reduced models B and MB are a suitable alternative to the F model, in terms of negligible residual confounding and increased precision and model fit.

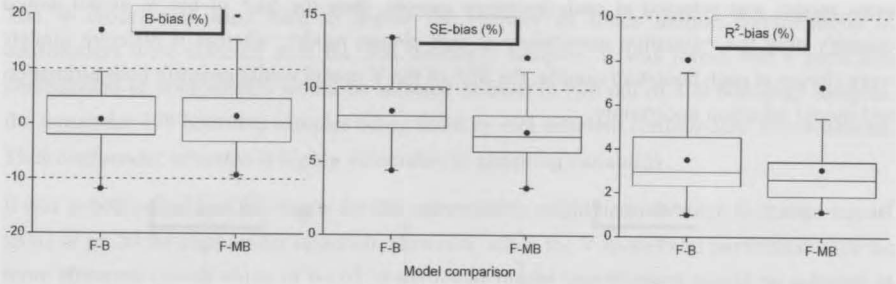



Figure 4.4 Boxplots (minimum, 25%, 50%, 75% and maximum), calculated over the 50 relationships, of the %bias between estimates obtained from the F model and the two reduced models: B and MB. Results are shown for the slope (B-bias), standard error (SE-bias) and multiple correlation coefficient ($R^2$).

Figure 4.4 shows that for most (90%) relationships, reduced models do not lead to substantial residual confounding (B-bias<|10%|).This result confirms the idea that variable selection with F-change at a large cut-off value, usually results in no substantial residual confounding. The fact that the B-bias is symmetric around zero implies that model reductions can lead to both positive and negative confounding (Tzelgov & Henik, 1991; MacKinnon et al, 2000; Friedman & Wall, 2005). Substantial gain in precision (SE-bias>10%) due to model reductions is observed in only 20% of the relationships. This again was expected owing to the large cut-off value for variable selection but may have been compounded by the strict criteria used to select the 50 relationships in the first place. Finally, the model's fit ($R^2$-bias) is not substantially changed by model reductions.

These results suggest that model reductions are probably not worthwhile given the variable selection criterion used. Since although the latter does not cause residual confounding, it also does not cause substantial gains in precision or model fit in the majority of the relationships. The only advantage of model reductions is that models are smaller and thus easier to interpret.

### 4.4.4 Model selection uncertainty

In the previous section, both confounder selection and estimation were performed on the same data. This violates an assumption of least squares estimation: that the data is collected to perform estimation conditional on a model that was either chosen a priori or chosen on independent data. Model selection from 30 or 33 confounders could theoretically deliver $2^{30}$=1 073 741 824 or $2^{33}$=8 589 934 592 possible models which implies huge levels of

multiple testing. Some methods to quantify model selection uncertainty a posteriori (i.e. after performing both model reduction and estimation on the same data) and even to adjust errors to more realistic levels have been suggested, but their effectiveness is still debatable (Chatfield, 1995; Faraway, 1992; Breiman, 1992).

Here, model selection uncertainty was quantified with a bootstrap-based method suggested by Chatfield (1995), which is referred as the V model here. The latter consists in performing both variable selection and estimation on each bootstrap model. It might be expected that if the same model was selected at each bootstrap sample, then the SE* of the V model would quantify only the estimation uncertainty of that chosen model; whereas if different models were chosen at each bootstrap sample, the SE* of the V model would quantify both estimation and model selection uncertainty.

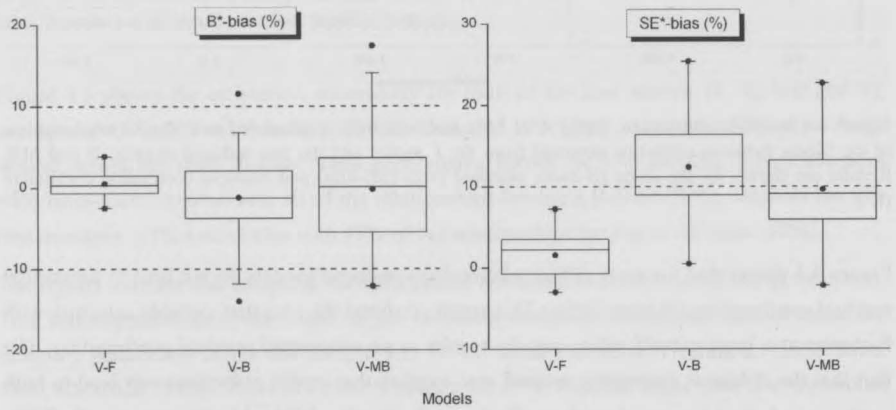



Figure 4.5 Boxplots (minimum, 25%, 50%, 75% and maximum), calculated over the 50 relationships, of the %bias between estimates between the bootstrap estimates obtained from the V model and those obtained from the other three models: F, B and MB.

Figure 4.5 shows that for nearly all relationships, the SE* of reduced models is lower than those of the V model. The difference is substantial (>10%) for about 52% of the relationships. This result could presumably be attributed to model selection uncertainty; however the uncertainty seems to be rather small. The Full model, which theoretically contains no model selection uncertainty, shows estimates that are remarkably similar to those of the V model.

As a method for quantifying model selection uncertainty, the V model seems to be inadequate, for two reasons. First, relationships where the reduced models did not significantly increase precision relative to the F model (approx. 80% of the relationships, Figure 4.4) have their model selection uncertainty un-quantified because the SE* of the reduced models is indistinguishable from that of the F model, which in turn is indistinguishable from that of the V model. For instance, the relationship between Cerebrovascular diseases in females 45-64 years of age and K shows identical SE* (and B*) regardless of which model is used (F, B, MB or V). Second, to the extent that estimates of the

V model are identical to those of the F model, it is more efficient and less computationally intensive to use the F model as the reference model for assessing model selection uncertainty.

### 4.4.5 Robustness of confounder selection

Robustness of confounder selection to sampling variability provides an indication of whether a true single model exists, a property which can then be used for reducing models on the basis of model averaging, as was performed here with the MB model.

The V model was used here to assess the number of times unique combinations of confounders were selected over the 500 bootstrap samples. It was found that a particular combination of confounders would be selected at most in two out of 500 bootstrap samples, the remainder 498 bootstrap samples being fitted by 498 different confounders' combinations. Thus confounder selection is highly vulnerable to sampling variability.

It was hypothesised that the reason for the vulnerability might lie in the high F-change cut-off value of $p<.20$ for confounder selection. However, when the V model was performed with the more stringent cut-off value of $p<.05$, a particular model specification would be selected at most in 22 out of 500 bootstrap samples. Then it was hypothesised that the reason might lie in the requirement that confounder selection be performed backwards, when collinearity is at it highest. Excessive collinearity is known to cause instability in the predictors' partial correlation coefficients, the latter being used for variable selection (Cohen et al, 2003; Friedman & Wall, 2005). To test this hypothesis, the V model was ran with forward selection with the F-change cut-off value of $p<.20$. In this case, a particular model specification was selected in no more than 80 out of 500 bootstrap samples.

Robustness of confounder selection was also assessed with respect to its consistency within and between hospitalisation categories. The inclusion frequency of confounders was highly (average $R^2>.80$) correlated within hospitalisation categories (i.e. same diagnostic-gender-age category, but with different pollutants), and was poorly correlated between hospitalisation categories (i.e. different diagnostic-gender-age category) (average $R^2>.10$). This suggests that the pollutant is not an important determinant of which confounders are retained in the model. Figure 4.6 presents an example of the correlation of the inclusion frequency of each confounder in two relationships predicting ischemic heart disease in females >64 years of age from As and Ni.

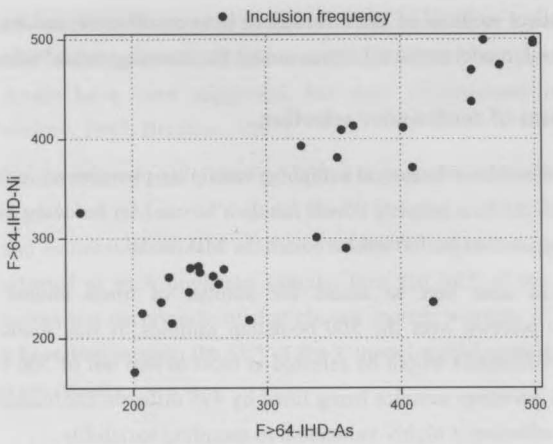



Figure 4.6 Inclusion frequency of each of 30 confounders in two relationships involving the same hospitalisation category: ischemic heart disease in females >64 years of age, and two chemical elements: Arsenic and Nickel. The inclusion frequency is the number of times each confounder is retained in the model across 500 bootstrap samples (V model).

## 4.5 Discussion

Previous studies using atmospheric biomonitoring data have been mostly exploratory, using correlation measures and lacking adequate control for confounding (Cislaghi & Nimis, 1997; Wappelhorst, 2000; Wolterbeek & Verburg, 2004a; Sarmento et al, 2008). This study is also exploratory but, to the best of our knowledge, it is the first to incorporate sounder epidemiological methods.

The interpretation of the associations obtained by this study as causal effect measures, hinges on the willingness to accept some of its pitfalls. First, hospitalisation data probably contained non-negligible repeated counts for the same patient, and thus it is of less aetiological value than pure incidence data (Rothman, 2002). Second, all variables were ecological and most could not be standardised by gender and age (mutual standardisation bias; Rosenbaum & Rubin, 1984). Third, owing to the absence of potentially important confounders (e.g.: lifestyle and physiological), and the fact that only single-pollutant models were investigated, effect measures are almost certainly biased upwards (Chen et al, 1999). Fourth, the study is of an ecological aggregate design and thus it is subject to several biases, collectively known as ecological or cross-level bias (Greenland & Morgenstern, 1989; Salway & Wakefield, 2004; Glynn et al, 2008; Firebaugh, 1978). Most of these biases cannot be detected and/or solved with ecological data alone. Fifth, it is unclear how trace metal elements in lichens might relate to those found in the atmosphere by instrumental monitoring (Wolterbeek, 2002). In addition, it is unclear to what extent levels of trace metal elements in the atmosphere or in atmospheric deposition may be related to levels in other media such as food and water. Sixth, the between-area to within-area variance ratio (B/W) in the trace metal elements concentrations over the study area (Table 4.2) was probably grossly over-estimated due to under-sampling of the

within-area variance (Wolterbeek et al, 2010; Wolterbeek & Verburg, 2002, 2004b; Salway, 2003 Chapter 3). This means that the magnitude of the bias from this particular source is likely to be upwards and large (Salway, 2003 Chapter 3; Wakefield, 2008; Greenland & Morgenstern, 1989; Firebaugh, 1978; Webster, 2007). This important problem is likely affecting similar studies, both multi-level and aggregate ecological, although studies rarely if ever report the B/W for air pollution exposure (e.g.: Lipfert et al, 2000, 2006; Pope, 1995; Dockery, 1993).

Over the 50 selected relations, the effect of the pollutants was always positive except for one relationship involving Cl. As was the most conspicuous element, being associated with 13 hospital admissions categories. Coincidentally or not, inorganic As is notorious for its cardiovascular effects following ingestion but the evidence is less convincing for inhalation exposures (ATSDR, 2007 p58). The As associations could be due to soil contamination, which could lead to higher levels of As in air and lichens, as well as in water and food. Many strong effects were observed for seemingly innocuous elements such as Mg, K and Fe, which are probably partly associated with lichen physiology, and thus are implausible. Ni and V have been shown to cause respiratory toxicity in vitro and in vivo (Ghio et al, 2002; Dye et al, 2001), and they have also been found to be the principal predictors of cancer in the Portuguese population (Sarmento et al, 2008) and of total mortality in the US Veterans cohort (Lipfert et al, 2006). Among the emission factors, F2, which is the main emission source for Ni and V, was the most conspicuous and had the strongest effect.

This study suggests that a 1% increase in pollutant is associated with an average 15% increase in hospitalisations (min-max: -15%-33%) (Table 4.6). Elasticities reported for chronic diseases and traditional air pollutants such as PM10 and Sulphates have been in the order of 5-15% (Lipfert & Wyzga, 1995; Lipfert et al, 2000; Lipfert, 1997). In this study the elasticity at the mean for Ni and V also averaged 15% (min-max: 10-28%) whereas in a follow-up of the Washington University Veterans cohort, the same metals showed elasticities of 5% and all other chemical elements showed non-significant effects (in single-pollutant models) (Lipfert et al, 2006).

In addition to reporting tentative effect estimates for trace metal elements, this study aimed at investigating issues of estimation and confounder selection. By tapping the bootstrap's analytical potential it was possible to investigate the robustness of effects estimates and confounder selection to sampling variability, and model selection uncertainty.

It was found that SEs in half of the relationships were substantially affected by sampling variability and this was likely due to data characteristics such as outliers. In 14 relationships using SE* instead of SE led to the pollutant's effect losing statistical significant at $p<.01$ (Table 4.6). Had SE* been used in the initial selection of significant relationships, these 14 relationships would not have been chosen. Thus it may be informative to routinely use SE* in selection of significant relationships and all ensuing analyses.

For the vast majority of the relationships, model reductions do not cause residual confounding but they also don't increase precision substantially. Thus model reductions are not

recommended. This conclusion confirms what has been found by others (e.g.: Jorgensen et al, 2007; Fewell et al, 2007). The requirement, in causal analyses, that variable selection be performed backwards and at a high F-change cut-off value in order to prevent residual confounding, obliterates the main advantage of using prediction-based model selection criteria in the first place: to enhance precision (Mickey & Greenland, 1989; Maldonado & Greenland, 1993; Greenland, 1989); furthermore any enhanced precision may be too optimistic due to model selection uncertainty, which leads to the next assessment.

It is a well-known, but often ignored, fact that variable selection and estimation should not be performed on the same data, because the uncertainty associated with multiple testing is not accounted for in the estimation. Some methods to quantify this model selection uncertainty a posteriori (i.e. after performing both model reduction and estimation on the same data) and even to adjust SE to more realistic levels have been suggested, but their effectiveness is still debatable (Chatfield, 1995; Breiman, 1992; Faraway, 1992). This study used a method suggested by Chatfield (1995), and which is called here the V model. It appears that this method does not convey any additional information concerning the model selection uncertainty in reduced models, than that provided by the F model where no variable selection was performed. Since the V model is highly computationally intensive it seems more efficient to use the F model for this purpose.

Confounder selection was found to be very vulnerable to sampling variability. This issue is important because it indicates whether a "true" model exists among the confounders. One of the reasons that might have contributed to this vulnerability is the high collinearity among confounders (Cohen et al, 2003; Friedman & Wall, 2005). For instance, the Tolerance (Table 4.6) of the pollutants averaged 0.62 (min-max: .46-.81), whereas the Tolerance of the confounders was much lower averaging 0.20 (min-max: .01-.71), over the confounders in all relationships. Excessive collinearity renders semi-partial correlations very unstable to small changes in parameters or data, leading to unstable confounder selection (Friedman & Wall, 2005; Tzelgov & Henik, 1991; Cohen et al, 2003). This implies that model averaging based on sampling variability, such as the MB model, may not be representative. Confounder selection was, however, quite consistent within hospitalisation categories. As a basis for model averaging, this consistency appears to be more representative of the persistency of substantial associations between confounders and health variables (e.g.: Jorgensen et al, 2007; Heymans et al, 2007).

## 4.6 Conclusions

The effect estimates for trace metal elements presented in this paper are surprisingly coherent with expectations, because effects were adverse for nearly all pollutants, many of which have been implicated in cardiovascular disease (e.g.: As, Ni and V), and because elasticities are in the range of those found for traditional air pollutants, if somewhat higher as might be expected from more proximal causal agents. However, these results must be interpreted with extreme caution because inhalation exposure to atmospheric pollution was measured indirectly through biomonitoring. Furthermore effect estimates may be severely inflated by at

least three inter-related factors: 1) omission of important between-area confounders, 2) confounding and effect modification by group in the exposure, and 3) a low true between-area to within-area variance ratio in the exposure of interest (Greenland & Morgenstern, 1989; Webster, 2007; Salway, 2003).

The non-parametric bootstrap, by creating sampling variability, is a useful tool to assess and correct the robustness of effect estimates and to assess the robustness of variable selection. However it does not appear to be a suitable or efficient means of assessing model selection uncertainty.

# 5 Suppression situations in the geographical association between trace metal elements, measured by atmospheric biomonitoring, and circulatory disease – application and implications for environmental epidemiology

*Based on article:*
*Sarmento SM., Verburg TG, FreitasMC & Wolterbeek HTh.*
*Submitted to Inhalation Toxicology,*
*January 2012.*

## 5.1 Abstract

Context: Confounding can be positive (redundancy) or negative (suppression). The distinction is important due to their antagonistic effects on slopes, errors and model fit: redundancy often leads to "pessimistic" results, whereas suppression often leads to "optimistic results". The distinction is especially important when the exposure of interest is a weak predictor and when extensive control for confounding is required, as is often the case in environmental epidemiology.

Objectives: First, demonstrate the application of recent methods of identification of confounding types which have been mostly confined to the Social Sciences. Second, observe how frequent suppression situations are in a real-world epidemiological dataset. Third, interpret confounding types in terms of causal mechanisms, and evaluate their impact on standard errors and model fit.

Materials & Methods: Aggregate ecological study compared hospitalisations due to cardiovascular diseases, with the concentration of selected air pollutants, measured through biomonitoring, across 125 municipalities in Portugal. Single-pollutant models with 30-33 confounders were estimated with linear regression. Identification of confounding types was based on modern definitions and criteria.

Results: Suppression situations are not uncommon and most relationships affected by it are difficult to reconcile with a causal mechanism; however their inflation of errors and model fit is negligible because they tend to show low collinearity levels.

Conclusion: It is recommended that identification of confounding situations should be performed routinely to screen what may appear, at first sight, large and statistically significant effects, for inconsistencies with a causal explanation and for volatile effect estimates and inflated statistical significance.

## 5.2 Introduction

Confounding is one of the most important biases in observational epidemiology (Wakefield, 2003; Hayes, 2003). In a sense, control for confounding is an experiment ran backwards. One observes a phenomenon in the wild, under uncontrolled conditions, and then imposes constraints on the resulting data, to mimic experimental conditions. These constrains are imposed mostly trough mathematical modelling, guided by substantive reasoning.

Confounding has been defined in many ways: counterfactuals, randomisation, comparability and collapsibility (Greenland & Morgenstern, 2001; Morgenstern, 2008; Salway & Wakefield, 2004; McNamee, 2003). The definitions with greatest practical relevance are comparability and collapsibility, which tend to be used in combination. The comparability definition provides a set of characteristics that a confounder (Z) should have (e.g.: causally associated with the outcome and non-causally correlated with the exposure of interest), and therefore it is used to select potential confounders (McNamee, 2003). The comparability definition may need to be extended when control for cross-level bias is required (e.g.: confounding by group; Morgenstern, 2008; Willis et al, 2003; Salway, 2003). However, this definition cannot confirm whether the potential confounder causes confounding on a particular dataset and model. For this objective, the collapsibility definition takes over. According to this definition, which is at the core of the Change in Estimate criterion for variable selection, the relationship between Y and X is confounded by a third variable Z, if inclusion of Z in the model substantially changes the effect of X. Thus, collapsibility compares the effect of the exposure of interest between two models: one with confounder(s) and another with no or just a subset of confounder(s). If the difference is large, confounding is deemed substantial and the confounder(s) should not be excluded from the model.

In observational epidemiology, confounding control is often, but perhaps unwittingly, conveyed as a conservative procedure: inclusion of confounders in the model tends to change the effect of the exposure of interest towards the null, and the consequent decrease in degrees of freedom and increase in collinearity, tend to increase errors and decrease model fit. This scenario, known as positive confounding or redundancy, however, is only one of several ways in which confounding can affect model parameters. Other scenarios exist, where confounding has quite the opposite consequences: increasing effect estimates and model fit, and decreasing errors (or some combination thereof). This latter scenario is generally known as negative confounding or suppression.

To have a better feeling for these two types of confounding situations, consider the following, admittedly contrived examples.

Example 1: A daily time-series study attempts to correlate non-infectious respiratory disease and outdoor SO2 concentrations. The incidence of influenza is a potential confounder because it is a risk factor for the development of non-infectious respiratory diseases and it is often positively associated with SO2. Failure to control for influenza would lead to inflated SO2 slopes, because the latter will express not just the unique effect of SO2 but also the effect of influenza through its correlation with SO2.

Example 2: Suppose now the same scenario, but the exposure of interest is O3 instead. O3 is usually negatively correlated with influenza incidence. Thus, lack of control for influenza will suppress/mask the O3 effect, because the latter will express both the unique effect of O3 minus the effect of influenza through its correlation with O3.

The two examples may be understood more fully by using a formula for the simple case of linear regression with three standardised (z-score) variables: the outcome (Y), the pollutant (X) and the potential confounder (Z). Note that the crude effect of X and Z ($r_{yx} > 0$ and $r_{yz} > 0$) are assumed positive, but collinearity can have either sign. The effect of X, adjusted for Z, is given by $b_{yx} = \frac{r_{yx} - r_{yz}r_{xz}}{1 - r_{xz}^2}$ (equation 1). This shows that, when collinearity is zero ( $r_{xz} = 0$), the association of X with Y remains intact, i.e. $b_{yx} = r_{yx}$, and thus no confounding bias exists. In the SO2 example, the collinearity was positive ($r_{xz} > 0$) and thus a possible (but not unique) outcome is $b_{yx} < r_{yx}$. This is redundancy, and it implies that influenza inflates the crude association of SO2. In the O3 example, the collinearity was negative ( $r_{xz} < 0$) and thus the only possible outcome is $b_{yx} > r_{yx}$. This is suppression, and it implies that influenza suppresses or masks the crude association of O3.

The suppression situation illustrated by the O3 example, although unusual, is logical since it is caused by negative collinearity. However, this is only one of the three types of suppression recognised so far (Friedman & Wall, 2005). The other two suppression situations, occur when collinearity is positive, and lead not only to $b_{yx}$ being larger than $r_{yx}$, but also of a different sign. One renowned cause for this is when collinearity is positive and large, and when X is a weaker predictor than Z. In this case, the numerator in the formula above is $r_{yx} < -r_{yz}r_{xz}$, leading to a negative adjusted association. Another, more obscure, cause for the suppression situation just described is when the crude association of X is very low, so that essentially only $r_{yx} < -r_{yz}r_{xz}$ can occur, regardless of collinearity levels (Friedman & Wall, 2005).

The main message of the description above is that all three parameters: $r_{yx}$, $r_{yz}$ and $r_{xz}$ conspire to bring about suppression situations, and although collinearity is essential, it is not the sole culprit. Suppression may arise at seemingly low collinearity levels when X is a weak crude predictor.

Investigation of the multiple ways in which confounding impacts effect estimates and other model parameters appears to be restricted to the Social Sciences and mediation analysis. Since the 1940s researchers have attempted to pin down the different types of confounding and to find reliable criteria to distinguish them (e.g.: MacKinnon et al, 2000; Tzelgov & Henik, 1985, 1991; Tzelgov & Stern, 1978; Velicer, 1978; Horst, 1941). One of the earliest trends in this research was the use of graphical displays to represent the different types of confounding as a function of all possible combinations of $r_{yx}$ and $r_{yz}$ in relation to $r_{xz}$. Research on this

puzzling topic was quite bumpy, with some literature containing incomplete or misleading definitions and graphical displays. However, it appears to have reached solid ground with the definitions, criteria and graphical display laid out by Friedman & Wall (2005), which unifies and improves on much of the work performed on the subject.

This paper's overarching aim is to demonstrate that the assessment of not just the extent, but also the type of confounding, affecting epidemiological relationships, can be instrumental to stimulate substantive reasoning and guide mathematical modelling. This statement is grounded on four rationalisations. First, identification of the type of confounding can offer some insight as to how the exposure of interest contributes towards the explanation/prediction of the outcome variable. For example, whether it is likely to be causal, whether it is suppressed, masked or redundant, and whether it is some sort of algebraic inevitability (Cohen et al, 2003; Maassen & Bakker, 2001; Tzelgov & Henik, 1991). Second, effect estimates in certain suppressed relationships are more likely to be found statistically significant, than those in redundant relationships, and thus are more likely to be reported (Friedman & Wall, 2005). Third, suppression situations are more common in circumstances that are typical of environmental epidemiology, namely: aggregate data, large models, and large collinearity levels (Cohen et al, 2003; Friedman & Wall, 2005; Morgenstern, 2008; Greenland & Morgenstern, 1989). Finally, suppression situations are more common when the exposure of interest is a weak predictor, regardless of collinearity levels (Cohen et al, 2003; Friedman & Wall, 2005). This situation is quite common in environmental epidemiology (Hayes, 2003; Wakefield, 2003), although authors rarely report crude associations (notable exception Lipfert et al, 2000).

The present paper has three specific aims.

First, to demonstrate the application of methods to distinguish between confounding types, in an epidemiological context. This is important because the subject appears to be restricted to the Social Sciences and mediation analysis. This aim is accomplished in the methods section, with deep reliance on the work developed by Friedman & Wall (2005) and Tzelgov & Henik (1991).

Second, to investigate how common are the different types of confounding in a real-world epidemiological dataset. Most discussions on the subject have used simulated datasets and explored how frequently confounding types could occur in theory over the full range of $r_{yx}$, $r_{yz}$ and $r_{xz}$. Instead the aim here is to known how frequently confounding types are likely to occur, in a real-world dataset, where relationships between variables are bounded by biological, social and economic contingencies.

The final aim is to interpret the confounding types in terms of causal mechanisms, and to evaluate how they affect model fit and standard errors, and thus statistical significance.

To accomplish these aims an aggregate ecological design was used, where the unit of analysis and the study area were 125 municipalities in Continental Portugal. Human exposure to airborne pollutants was assessed indirectly by lichen biomonitoring, and indicators included

32 chemical elements and eight emission factors. The Full model, which contained a single pollutant and all confounders selected a priori on the basis of substantive reasoning, was contrasted with a similar model with zero collinearity, in terms of several model parameters.

Considering the large number and diverse origins of the pollutants being examined, it might be expected that, assuming they are all detrimental to human health, some of them would correlate positively with the confounders and thus create a redundancy situation, whereas others would correlate negatively with the confounders, and thus create suppression situation. Still other pollutants may show high collinearity and/or low crude associations, which could also lead to suppression.

## 5.3 Methods

### 5.3.1 Hospital admissions database

See section 4.3.1 of this thesis.

### 5.3.2 Trace metal elements database

Atmospheric exposure to chemical elements was assessed indirectly by biomonitoring with lichens, which reflects the composition of atmospheric deposition both from atmospheric suspension and local re-suspension sources. Existing studies suggest that the correlation between biomonitors and instrumental measurements tend to correlate moderately well in most cases (reviewed by Wolterbeek, 2002).

The concentration of chemical elements was obtained from a biomonitoring survey that sampled the lichen *Parmelia sulcata* in the summer of 1993, throughout the territory of Continental Portugal. The database contained the concentration ($\mu g\ g^{-1}$ lichen) of 32 chemical elements in 228 sampling sites (black squares, Figure 4.2). Concentrations were determined by multi-elemental nuclear techniques: $k_0$-INAA and PIXE. A more detailed account of the sampling and analytical procedures may be found in Reis (2001), Reis et al (1996) and Freitas et al (1997, 1999, 2000).

This chemical element database was processed by Monte Carlo Target Transform Factor Analysis (MCTTFA), which identified eight emission sources (Kuik, Blaauw et al, 1993; Kuik, Sloof & Wolterbeek, 1993; Kuik & Wolterbeek, 1995).

Table 5.1 shows descriptive statistics for a selection of the 32 chemical elements and eight emission factors found to be significant predictors of hospital admissions in single-pollutant models. Of the eight emission factors identified, four (F1, F2, F3 and F5) were found to be significant predictors of hospital admissions. The emission factor F1 appears to indicate a soil source since it contributes to a large fraction of the occurrence (approx. 30%) of a wide number of soil-related elements: Sc, Fe, Ti, Th and Sm, and it tends to concentrate in the mostly rural east. F2 is associated with a fuel combustion source, since it contributes greatly to the occurrence of Ni and V (approx. 50%) followed by I, Pb and Sb (approx. 30%) and its geographical distribution is consistent with urban and industrial locations. F3 appears to

indicate a very localised soil source, characterised by the occurrence (factor contribution approx. 30-40%) to U, Rb, Cu, Cs and Th with high values in the north-east. F5 appears to be a mixed factor, associated partly with a sea source and partly with an As source. It contributes substantially towards the occurrence of just three elements: Cl, Na and As (approx. 45%). Its geographical distribution is fairly homogeneous along the coast, consistent with a sea source, with some hotspots in the interior, possibly associated with As-rich soils or with the use of As-based pesticides in vineyards (Freitas et al, 1999, 2000).

Table 5.1 Descriptive statistics (mean, standard deviation, minimum and maximum) of the concentration ($\mu g\ g^{-1}$ lichen) of chemical elements and their associated emission factors (prefix F) determined in the lichen *Parmelia sulcata* in 227 sampling sites, corresponding to 125 municipalities in Continental Portugal (Figure 4.2). Only those chemical elements and emission factors that were found to be significant predictors of hospital admissions are shown.

| | Mean | SD | Min | Max | | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|---|---|---|
| Al | 5383 | 2494 | 1940 | 13 400 | Ni | 3.76 | 2.06 | 1.33 | 10.60 |
| As | 1.72 | 0.93 | 0.71 | 4.85 | Rb | 16 | 7.5 | 5.5 | 41 |
| Cl | 1365 | 481 | 528 | 3200 | Sb | 0.29 | 0.16 | 0.11 | 0.84 |
| Cr | 5.26 | 2.28 | 1.96 | 13.40 | Sc | 0.40 | 0.14 | 0.16 | 1.02 |
| Cs | 0.60 | 0.30 | 0.22 | 1.72 | Sm | 0.44 | 0.20 | 0.15 | 1.16 |
| Eu | 0.18 | 0.08 | 0.07 | 0.48 | Th | 0.88 | 0.46 | 0.32 | 2.48 |
| Fe | 2126 | 989 | 705 | 5320 | Ti | 330 | 153 | 111 | 808 |
| Hf | 0.41 | 0.20 | 0.14 | 1.12 | U | 0.25 | 0.13 | 0.10 | 0.69 |
| I | 6.78 | 3.18 | 2.24 | 17.60 | V | 14 | 7.87 | 5.35 | 41 |
| K | 5463 | 1562 | 2280 | 10 900 | F1 | 62 | 65 | 0.00 | 362 |
| La | 2.98 | 1.46 | 1.01 | 7.80 | F2 | 10 | 8.21 | 0.00 | 47 |
| Mg | 1987 | 742 | 772 | 4690 | F3 | 23 | 18 | 0.00 | 98 |
| Mn | 51 | 18 | 19 | 115 | F5 | 39 | 19 | 9 | 110 |

### 5.3.3 Confounders database

See section 4.3.3 of this thesis.

### 5.3.4 Study area and unit of analysis

See section 4.3.4 of this thesis.

### 5.3.5 Selection of relationships

The databases, consisting of 16 diagnostic-gender-age hospital admission categories and 40 pollutants resulted in no less than 640 possible single-pollutant relationships.

Selection of just the most significant relationships was performed using a criterion that is typical of environmental epidemiology (e.g.: Lipfert et al, 2000; Jorgensen et al, 2007). This involves estimating the pollutant's effect in a model which contains all confounders selected a priori, possibly performing some variable selection (e.g.: Change in Estimate criterion) to

improve precision, and then select those relationships where the pollutant's effect and model fit are both significant at some statistical level (usually p<.01 or .05).

Accordingly, the present work selected relationships where the pollutant's effect and model fit, estimated by the Full model which contained all confounders selected a priori, were both significant at p<.01. This yielded 67 relationships.

Identification of confounding types shall be performed only on these 67 relationships, since these would be the ones that would normally be reported in epidemiological studies.

### 5.3.6 Software

SPSS 17.0 syntax was used to perform Ordinary Least Squares linear regression, and MS Excel 2003 for all other calculations. ArcGIS Explorer Desktop was used to plot the geo-referenced map in Figure 4.2.

### 5.3.7 The collapsibility definition of confounding

The collapsibility definition of confounding, described in the introduction, requires two decisions from the researcher.

First, two model specifications must be chosen for comparison. Usually one of the models is the Full model, which includes the pollutant and all confounders, whereas the other is some reduced model that includes the pollutant and excludes all or a subset of the confounders. Unquestionably, the extent and type of confounding depend on what models are compared (Tzelgov & Henik, 1991). For the present work, it was decided to compare the Full model, which contained a single pollutant and all 30-33 confounders, with the Simple model which equals the Full model except that the collinearity is set at zero. With the Simple model: $r_{yx} = b_{yx}$, $r_{yz} = b_{yz}$ and $R^2 = r_{yx}^2 + r_{yz}^2$. The operationalization of the Simple model as a Full model with zero collinearity, as opposed to a model which does not include confounders, makes it easier to compare other parameters besides the association of the predictors with the outcome, such as $R^2$ and standard error (SE).

Second, an effect estimate must be chosen to compare the pollutant effect in the two models. This depends on the type of model selected (e.g.: risk difference for linear model, odds ratio for logistic models). Incidentally, the choice of effect estimate has an impact on the extent of confounding, and possibly also on the type of confounding, but this latter issue needs more research (e.g.: Lynn, 2003). For the present work, linear effect estimates shall be used, for four reasons. First, the methods to identify types of confounding situations are well developed for linear models (Friedman & Wall, 2005) but not for linear models (e.g.: Lynn, 2003). Second, all variables are continuous and the study design is aggregate ecological; in such cases, linear regression is recommended (Rothman, 2002; Greenland, 1992; Greenland & Robins, 1994; Salway, 2003; Glynn et al, 2008). Third, the study area is Portugal, a fairly small and un-industrialised country which benefits from favourable dominant Atlantic winds. Thus it is reasonable to assume that exposure is low and has a narrow range, relative to the full exposure range of the true dose-response curve. In such cases, a linear approximation is

reasonable (Rothman, 2002; Wakefield, 2003; Salway & Wakefield, 2004). Fourth, non-linear models assume that the exposure effect interacts with the confounder, which seems unreasonable, especially in an ecological study (Rothman, 2002).

Identification of confounding types has used either beta-weights (Conger's definition) or semi-partial correlations (Velicer's definition) as the measure of the adjusted effect of the exposure of interest. For the present work, Conger's definition was chosen because: 1) it is more closely related to the epidemiological measure, the unstandardized slope; 2) it is the most consensual for the purpose of confounding identification; and 3) it has been shown to apply to linear combinations of variables (see next section) (Cohen et al, 2003; Friedman & Wall, 2005; Tzelgov & Henik, 1991).

### 5.3.8 Calculations to identify types of confounding

Identification of confounding types, as laid out by Friedman & Wall (2005), require the comparison of three estimates obtained from the Full model and the Simple model: association of the pollutant with the outcome, association of the confounder(s) with the outcome, and the multiple correlation coefficient ($R^2$). In some situations, only a fraction of these three parameters are necessary to recognise the type of confounding. Nevertheless, for wider applicability, all three were estimated. In addition, it is instructive to calculate the collinearity in the Full model, because it is an indispensable parameter in determining the extent and type of confounding.

The calculations necessary to identify confounding types are listed below.

Calculations have been developed for the trivariate case only, i.e. one outcome (Y), one exposure (X) and one confounder (Z). The Full model, however, contains a single pollutant and 30-33 confounders. The linear combination method has been suggested for aggregating a functional set of variables into a single one (Friedman & Wall, 2005; Tzelgov & Henik, 1985, 1991; Holling, 1983; Cohen et al, 2003). Following this method, the linear combination of the 30-33 confounders (denoted by Z*) was calculated by saving the standardised predicted values from the linear regression of the hospital admissions on the 30-33 confounders: $Y = b_{yz1}Z_1 + b_{yz2}Z_2 + ... + b_{yz33}Z_{33}$.

When necessary, the pollutant variables were oriented (i.e. multiplied by -1) so that their zero-order correlation with the hospitalisations was always positive ($r_{yx} > 0$). By definition, the zero-order correlation for a linear combination of variables is always positive ($r_{yz*} > 0$). This procedure is only meant to simplify interpretation because it ensures that only $r_{xz}$ can be negative in equation 1 (Friedman & Wall, 2005; Tzelgov & Henik, 1991). With our relationships, 6 out of 67 relationships had negative zero-order correlations and thus were oriented. Out of these 6 relationships, 5 saw their beta-weight reverse sign relative to the crude correlation (Figure 5.2 - NRS).

Estimated the zero-order correlation of the hospital admissions with the exposure ($r_{yx}$) and with the linear combination of confounders ($r_{yz*}$). This is the same as the beta-weight provided in the Simple model, which is specified with the pollutant and confounders, but zero collinearity.

Estimated the collinearity ($r_{xz*}$) by regressing the pollutants on the linear combination of confounders: $X = r_{xz*}Z^*$.

Estimated the beta-weight of the pollutants ($b_{yx}$) and the beta-weight of the linear combination of confounders ($b_{yz*}$) from the Full model specified with the linear combination of confounders: $Y = b_{yx}X + b_{yz*}Z^*$.

Estimated the multiple correlation coefficient of the Full model ($R_F^2$), specified with the linear combination of confounders, and the of the Simple model where collinearity is zero: $R_S^2 = r_{yx}^2 + r_{yz}^2$.

It is important to note that other model parameters, such as standard errors (SE) and t-values are also affected by the different types of confounding. However, their determination is not essential for the identification of confounding types. Our aim, however, includes an assessment of how the different confounding types affect SEs, and thus statistical significance. Therefore, for each of the 67 relationships, the SE of the Full model ( $SE_F\left(b_{yx}\right) = \sqrt{\frac{1-R_F^2}{n-k}\frac{1}{1-r_{xz}^2}}$ ) was compared with that obtained from the Simple model where collinearity is zero: $SE_S\left(b_{yx}\right) = \sqrt{\frac{1-R_S^2}{n-k}}$, where $R_S^2 = r_{yx}^2 + r_{yz}^2$. To facilitate comparisons over the 67 relationships $n-k$ was set at 210.

### 5.3.9 Introduction to the types of confounding

This section describes the four types of confounding situations recognised so far, and explains how the difference between the crude and adjusted association between the exposure and the outcome, the crude and adjusted association between the confounders and the outcome, and the $R_S^2$ and $R_F^2$, calculated in the previous section, can be used to identify them. This description applies only to trivariate linear models with standardised variables and is based on the work of Friedman & Wall (2005).

Note that it is assumed throughout that $r_{yx} > 0$ and $r_{yz} > 0$ because, as it will become apparent this simplifies interpretation. It is also assumed that $r_{yx} < r_{yz}$, i.e. that the exposure of interest is a weaker predictor than the confounder(s), since this is the most common

situation in environmental epidemiology (Wakefield, 2003; Hayes, 2003). As shall be seen, the extent and type of confounding affect the weakest predictor more strongly.

Figure 5.1 depicts the four types of confounding (4 columns) and their main characteristics (3 rows), in terms of the change they exert on the association estimates of X and Z with the outcome (Y) and on model fit ($R^2$). The four types of confounding follow a continuum over collinearity levels, so that for instance, Positive Reciprocal Suppression (PRS) only occurs when $r_{xz} < 0$, whereas Redundancy only occurs when $r_{xz} > 0$ and is not too extreme. Within the interval of $r_{xz}$ that each confounding type occupies, the change in parameters, i.e. the confounding bias, tends to become more pronounced as (absolute) collinearity increases.

Figure 5.1 is not completely accurate because it gives the impression that each confounding situation occupies an equal interval over $r_{xz}$. This is usually not the case, and in fact, for some values $r_{yx}$ and $r_{yz}$, some types of confounding may not be possible at all, regardless of collinearity levels (Friedman & Wall, 2005). However, the figure does summarise the main characteristics of the four confounding types, which shall be described in more detail below.

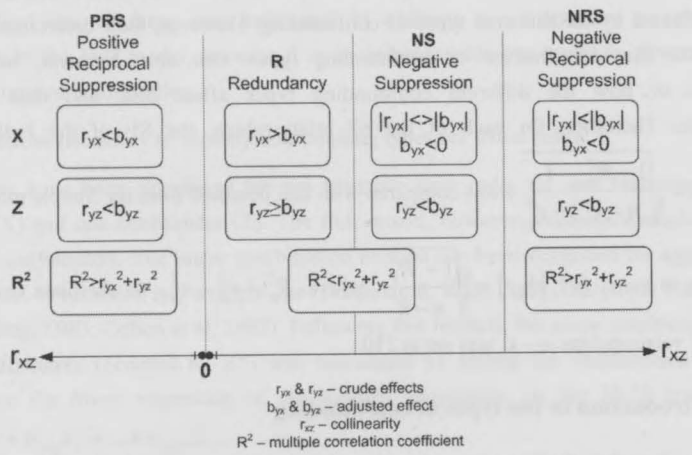



Figure 5.1 The four confounding types (columns) and the changes they exert on three model parameters (rows): association of the exposure (X) with the outcome, association of the confounder (Z) with the outcome, and multiple correlation coefficient (R2). Note that when collinearity is zero ($r_{xz}$): $r_{yx} = b_{yx}$, $r_{yz} = b_{yz}$ and $R^2 = r_{yx}^2 + r_{yz}^2$. Any departure from zero collinearity creates inequalities in these three expressions, which lead to the four types of confounding. Assumptions: $r_{yx} > 0$, $r_{yz} > 0$ and $r_{yx} < r_{yz}$. Note that for given values of $r_{yx}$ and $r_{yz}$, not all values of $r_{xz}$ are possible, and not all confounding types are possible or equally likely to occur over the possible range of $r_{xz}$.

Two general types of confounding exist: positive confounding (henceforth called redundancy) and negative confounding (henceforth called suppression). They are easily distinguished: $r_{yx} > b_{yx}$ is redundancy, whereas $r_{yx} < b_{yx}$ or $b_{yx} < 0$ is suppression.

There is only one type of redundancy, and it is reciprocal because both the exposure and confounder show a decrease in beta-weight relative to their respective zero-order correlations, i.e. $r_{yx} > b_{yx}$ and $r_{yz} > b_{yz}$, and it can only occur when $r_{xz} > 0$ (Figure 5.1).

Suppression, however, can be of three types: two are reciprocal in that both exposure and confounder show increases in (absolute) beta-weight relative to their zero-order correlations and the third type is mixed. Suppression situations can arise both when collinearity is positive or negative (Figure 5.1).

The most intuitive type of suppression is positive reciprocal suppression (PRS), because it is the mirror image of redundancy in interpretation and estimate changes. Here, $r_{yx} < b_{yx}$ and have the same sign (hence the name positive). It is reciprocal because the same happens to the confounder: $r_{yz} < b_{yz}$. PRS is the only type of confounding, when $r_{xz} < 0$ (Figure 5.1).

The second type of suppression is negative suppression (NS), which as the name implies, is not strictly reciprocal and it results in a negative beta-weight (hence the name negative) of the weakest predictor. So, $b_{yx} < 0$ and $r_{yx} > |b_{yx}|$, although as collinearity increases it becomes $r_{yx} < |b_{yx}|$. The confounder effect is also increased but does not reverse sign: $r_{yz} < b_{yz}$. NS only occurs when $r_{xz} > 0$ and is not too extreme (Figure 5.1).

Finally, the third and less intuitive type of suppression is called negative reciprocal suppression (NRS). As the name implies, it involves a change in the sign of the beta-weight of the weakest predictor (hence the name negative). Thus, $b_{yx} < 0$ and $r_{yx} < |b_{yx}|$, whereas $r_{yz} < b_{yz}$. NRS only occurs when $r_{xz} > 0$ and is reaching its limiting values (Figure 5.1).

Besides their different impact on the predictor's association with the outcome, confounding situations also impact model fit ($R^2$) differently. This is particularly useful to distinguish between NS and NRS situations. Under no collinearity: $R^2 = r_{yx}^2 + r_{yz}^2$. In PRS and NRS situations, the estimated $R^2$ is larger than this theoretical value and it tends to increase as collinearity increases, a phenomenon known as enhancement. Whereas in redundancy and NS situations, the estimated $R^2$ is smaller than this theoretical value and it tends to decrease as collinearity increases.

One of the interesting properties of confounding situations that show enhancement (i.e. NRS and PRS) is that standard errors (SE) *decrease* as collinearity increases. This contrasts with the more familiar situation where SEs increase as collinearity increases, observed in redundancy and NS situations (Friedman & Wall, 2005).

The interested reader is referred to the following publications for more detailed and comprehensive descriptions of confounding types: Friedman & Wall (2005), Tzelgov & Henik (1991), Maassen & Bakker (2001) and Cohen et al (2003).

## 5.4 Results

### 5.4.1 Identification of confounding types

The aim is to investigate the frequency with which each of the four confounding types arise in the 67 selected relationships, relating single-pollutants with hospitalisations in a fraction of the Portuguese population.

Identification of the confounding types affecting each relationship was performed by comparing three parameters (association of the pollutant with hospitalisations, association of the linear combination of confounders with hospitalisations, and the multiple correlation coefficient) between the Full model (i.e. pollutant and 30-33 confounders aggregated in a linear combination), and the Simple model (i.e. pollutant and 30-33 confounders aggregated in a linear combination but with zero collinearity).

Figures 5.2-5.4 show the three comparisons as a function of the collinearity estimated by the Full model, for each of the 67 relationships. Out of the 67 relationships, 44 show redundancy (R), 18 show PRS and 5 show NRS. No NS situations were detected, likely due to the strict criterion ($p<.01$ in Full model) used to select the 67 relationships. Under an NS situation, the pollutant generally presents a low (absolute) beta-weight and there is no enhancement (Figure 5.1). As a result, relationships under an NS situation present low adjusted associations with the outcome and are less likely to be found statistically significant, just as many relationships under redundancy were certainly not selected for the same reason.

Figure 5.2 shows the crude (white circles) and the adjusted (black circles) association of pollutants and hospitalisations, corresponding to the Simple and Full model estimates respectively, for each of the 67 relationships, as a function of collinearity. The distance between the two types of circles (vertical lines) is the confounding bias. As expected, the more collinearity diverges from zero, the greater the confounding effect. However, the direction of the confounding bias differs, giving rise to the different types of confounding.

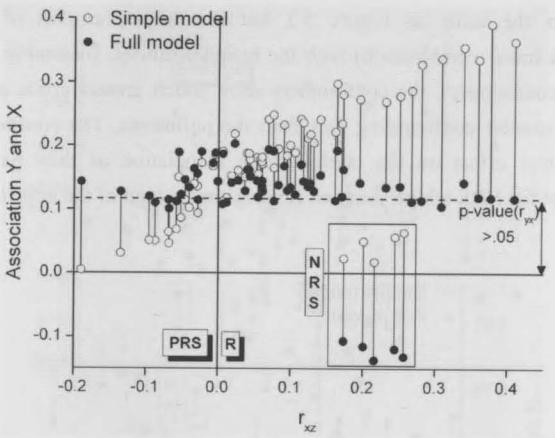



Figure 5.2 Crude (white circles) and adjusted (black circles) association of pollutant and hospitalisation in each of the 67 relationships, as a function of collinearity (x-axis). The line connecting the white and black circles is the confounding bias. The dashed horizontal line denotes the 5% threshold for statistical significance of the crude association. R-redundancy, PRS-positive reciprocal suppression, NRS-negative reciprocal suppression (encased by rectangles).

The five relationships under NRS (encased by rectangles) seem out of place in Figure 5.2 because collinearity levels, while positive, are not particularly high. Instead, this confounding situation can be attributed to the fact that the pollutant's crude association is very low and statistically not significant at $p<.05$ (white circles below the dashed horizontal line). In addition to these 5 relationships, 12 other relationships show crude associations that are very low and non-significant at $p<.05$. These relationships are the first 12 counting from the y-axis and are classified as PRS. What distinguishes these 12 relationships from the 5 relationships in NRS is the sign of collinearity. Owing to the very low (near zero) crude association of these 17 (5+12) relationships, it seems likely that its sign is arbitrary, which implies that the sign of their collinearity may easily reverse, and thus the fact they are classified as NRS or PRS is probably due to chance.

The relationships at the boundary of PRS and redundancy (R), on the other hand, have very low collinearity levels (i.e. $r_{xz} \cong 0$). Again the sign of this low (near-zero) collinearity is probably arbitrary and thus the fact that these relationships are classified as PRS or redundancy is also likely due to chance.

One of the most striking features of Figure 5.2 is that despite the fairly wide range of crude associations (.004-.391) over the 67 relationships, the range of beta-weights is much smaller (.101-.203). Furthermore, the higher the crude association of the pollutant the higher its collinearity.

Figure 5.3 shows the same as Figure 5.2 but for the association of the confounders (aggregated into a linear combination) with the hospitalisations. Inevitably if only due to the large number of confounders, the confounders show much greater crude associations (.591-.807), and much smaller confounding bias than the pollutants. The confounding types have the same qualitative effect on the confounder's association as they have on pollutant's association, except for NRS which show no reversal in the sign of the adjusted association.

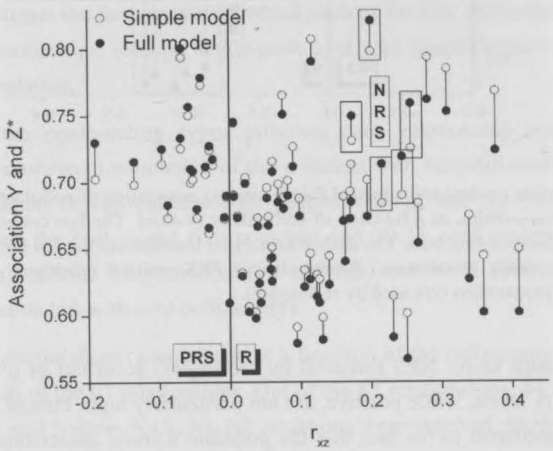



Figure 5.3 Crude (white circles) and adjusted (black circles) associations of the confounders (as a linear combination) in each of the 67 relationships as a function of collinearity (x-axis). The line connecting the white and black circles is the confounding bias. R-redundancy, PRS-positive reciprocal suppression, NRS-negative reciprocal suppression (encased by rectangles).

The comparison of Figure 5.3 and Figure 5.2 highlight the fact that the weakest crude predictor is more heavily affected by confounding bias and it is the one that may show a reversal in the sign of the adjusted association.

Figure 5.4 shows the same as Figure 5.3 and Figure 5.2 but for the multiple correlation coefficient ($R^2$). The black circles represent the $R_F^2$ of the Full model, whereas the white circles represent the $R_S^2$ of the Simple model when collinearity is zero (i.e. $R_S^2 = r_{yx}^2 + r_{yz}^2$). Again there is some tendency for the difference between the two to increase as collinearity increases; however, the direction of the difference varies. In redundancy, collinearity leads to smaller $R^2$, whereas in PRS and NRS, collinearity leads to larger $R^2$, the latter phenomenon being known as enhancement (Friedman & Wall, 2005). However, in PRS and NRS situations the increase in $R^2$ is very small and does not appear to increase with collinearity as rapidly as in redundancy situations.

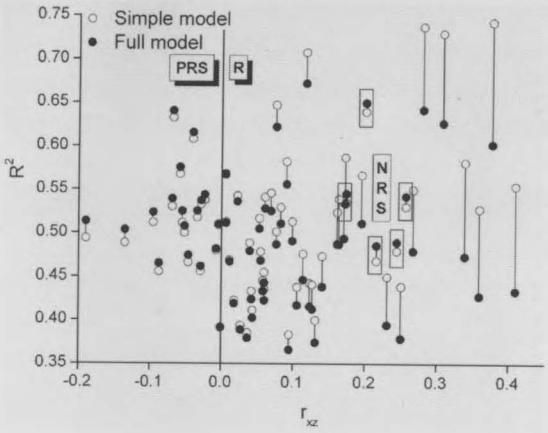



Figure 5.4 Multiple correlation coefficient ($R^2$) estimated from the Full model ( $R_F^2$, black circles) and from the Simple model where collinearity is zero ( $R_S^2$, white circles). R-redundancy, PRS-positive reciprocal suppression, NRS-negative reciprocal suppression (encased by rectangles).

### 5.4.2 Interpretation of confounding types

The 44 relationships found in redundancy situations are fairly easy to interpret. The crude association of the pollutants is not too low (>.15) and it is significant (p>.05), but is partly inflated by its positive correlation with the confounders (Figure 5.2). The 6 relationships under PRS which have fairly large pollutant crude associations (p<.05), have their effect partly masked by the pollutant's negative correlation with the confounders (Figure 5.2).

The relationships that challenge interpretation are the 5 under NRS and the12 under PRS that have non-significant crude associations (p>.05) (Figure 5.2). How to interpret relationships where the pollutant has a crude association that is indistinguishable from zero and an adjusted association that is large (in absolute terms) and statistically significant? There appear to be two possible interpretations, one is congruent with a causal mechanism, the other is not. One interpretation is that the crude association of the pollutant was completely masked by the confounders, and thus the crude effect was nearly zero. In the case of the 5 relationships under NRS, the original un-oriented data, had crude effects that were negative, which after control for confounding acquired the expected sign, positive. This observation may serve to reinforce our belief that these relationships are causal. An alternative interpretation is that the crude association of the pollutant is un-confounded, and its adjusted association is large because the pollutant explains error variance in the confounders (i.e. that part of the variance of the confounders that is not correlated with hospitalisations) (Cohen et al, 2003; Tzelgov & Henik, 1991; Maassen & Bakker, 2001). This interpretation suggests that the pollutant has no causal effect on the outcome variable.

The relationships at the boundary of PRS and redundancy (R) have very low collinearity, and thus very small changes in all parameters (Figures 5.2-5.4). Because confounding is regarded

as a deterministic, rather than probabilistic bias, a threshold of 10-15% change in the association of the exposure of interest has been recommended as a sign of substantial bias (Rothman, 2002; Robins & Morgenstern, 1987; Jorgensen et al, 2007; Fewell et al, 2007; Preacher & Hayes, 2008). Nevertheless, several relationships appear to not be substantially confounded, whichever cut-off criterion is used (Figure 5.2). The implication is that, for these relationships, the association of the pollutant is equally well estimated by a Full model as by a model with no confounders, because collinearity is basically nil. This finding is somewhat uneasy for, though possible, it is unlikely that these relationships are not confounded (Pearl, 1998).

### 5.4.3 The effect of confounding types on statistical significance

As seen previously (Figure 5.4) relationships in PRS and NRS situations show enhancement. Since $R^2$ is used in the calculation of SE, enhancement could cause SE to decrease (Friedman & Wall, 2005).

It is common knowledge that collinearity inflates SEs, however it does so only under a redundancy or NS confounding situation. Under PRS and NRS situations, collinearity deflates SEs. Many techniques have been developed to decrease SEs to more reasonable levels (e.g.: variable selection, ridge regression; Pitard & Viel, 1997; Cohen et al, 2003 Chapter 10; Goldberger, 1991 Chapter 23), however none appears to deal with the issue of too optimistic SEs.

Figure 5.5 shows the comparison between the estimated SE and the theoretical SE for the pollutant's association in each of the 67 relationships, as a function of collinearity. It is observed that, as expected, pollutants under a redundancy situation show increases in SE, which tend to become more pronounced as collinearity increases. Pollutants under PRS and NRS, on the other hand, show decreases in SE, which tend to be small and quite invariant to collinearity. A similar pattern is observed for the t-values, except that they decrease in redundancy and increase in PRS and NRS (not shown).

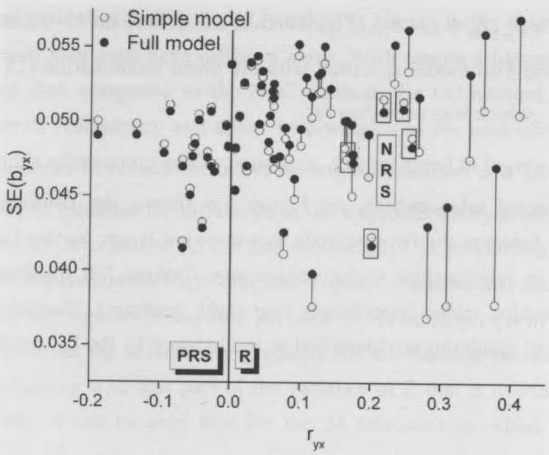



Figure 5.5 SE of the pollutant's association with hospitalisation, estimated from the Full model ( $SE_F$, black circles) and from the Simple model where collinearity is zero ( $SE_S$, white circles). R-redundancy, PRS-positive reciprocal suppression, NRS-negative reciprocal suppression (encased by rectangles).

### 5.4.4 A note on the use of the linear combination method to identify confounding types in the multivariate case

Before stating conclusions, a note is felt to be required with regards to one aspect of the methods used to identify confounding types: the linear combination method. All discussions of confounding types are restricted to the trivariate case (Friedman & Wall, 2005; Holling, 1983; Tzelgov & Henik, 1985, 1991; Cohen et al, 2003). This is due to difficulties in computation and interpretation, but probably also because the exposure of interest is usually assumed to be a stronger crude predictor than the confounder(s). In this case, it is not possible to identify all confounding types without knowing also the adjusted association of the confounder (Figure 5.1). In order to obtain an overall adjusted association for multiple confounders, the linear combination must be used.

However, when the exposure of interest is a weaker crude predictor than the confounders, as with the present dataset (Figure 5.2 and Figure 5.3), most confounding types can, in principle, be identified by just observing the confounding effect of the exposure of interest and thus it is not necessary to determine the adjusted association of the confounders through the linear combination method. As inspection of Figure 5.1 makes it clear, if the weakest predictor shows: 1) $r_{yx} > b_{yx}$ there is redundancy; 2) $b_{yx} > 0$ and $r_{yx} < b_{yx}$, there is PRS; 3) $r_{yx} > |-b_{yx}|$ there is NS; and 4) $r_{yx} < |-b_{yx}|$ there could be either NS or NRS depending on whether there is enhancement or not.

Thus, it should be possible to identify all confounding types by just observing the beta-weight of the pollutants from the Full model specified with all individual confounders:

$Y = b_{yx}X + b_{yz1}Z_1 + b_{yz2}Z_2 + ... + b_{yz33}Z_{33}$, which is called here the "No-Linear-Combination" (NLC) instead of the Full model specified with the linear combination (LC) of confounders: $Y = b_{yx}X + b_{yz*}Z^*$, which was used thus far.

It was found that the two methods disagree in the identification of confounding types for 24 out of the 67 selected relationships. As Figure 5.6 shows, the confounding bias though linearly correlated between the two methods, is somewhat larger for the LC method than for the NLC method in relationships under redundancy (bottom left quadrant) and somewhat smaller in relationships under suppression (top right quadrant). Twenty-four relationships (dashed bottom right quadrant) are identified as redundancy by the LC method and as PRS by the NLC method.

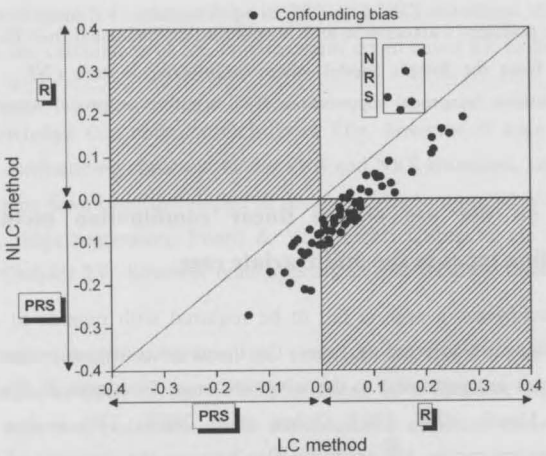



Figure 5.6 Confounding effect on the pollutant of each of the 67 selected relationships as estimated from the No-Linear-Combination (NLC) method and from the Linear Combination method (LC) method. Dashed areas correspond to when the two methods in the confounding type identified. R-redundancy, PRS-positive reciprocal suppression, NRS-negative reciprocal suppression (encased by rectangles).

The disagreement stems from the way collinearity is calculated by the two methods. With the LC method, collinearity is calculated only for that part of the confounders that is correlated with the outcome variable, i.e. the linear combination of confounders. Whereas with the NLC method, collinearity is calculated in the usual way, irrespective of the predictor's correlation with the outcome variable. In addition, while the collinearity calculated with the LC method is bounded by the crude association of the two predictors with the outcome, through the equation: $r_{xz} = r_{yx}r_{yz} \pm \sqrt{\left(1-r_{yx}^2\right)\left(1-r_{yz}^2\right)}$, which ensures a non-negative definite correlation matrix (Friedman & Wall, 2005); the collinearity calculated with the NLC method is not obliged to fall within this interval.

As a result, collinearity calculated by the LC method is always larger than that calculated by the NLC method, and may even have different signs. With simple arithmetics using equation 1, it is easy to see that, compared to the NLC method, the LC method gives larger beta-weights in the case of redundancy and lower beta-weights in the case of suppression, when the two method yield collinearities of the same sign, as confirmed in Figure 5.6.

Perhaps a better way to visualise the difference in the way collinearity is calculated by the two methods is to consider correlations (crude and collinearity) as proportions of variance, and schematise them with a Ballantine/Venn diagram. Figure 5.7 shows that the collinearity given by the NLC method corresponds to the total variance of X overlapping with the total variance of Z (area C+E), whereas the collinearity given by the LC method corresponds to the total variance of X overlapping with that part of the variance of Z that is overlapping Y (area C). With this framework, it can be seen that for the 24 relationships where the two methods disagree in the type of confounding, area C+E is negative whereas area C is positive. Although variances cannot be negative and the Ballantine/Venn diagram is probably not a reliable method to examine the problem (see "area C problem" in Cohen et al, 2003), it is shown here to help visualise this curious problem.

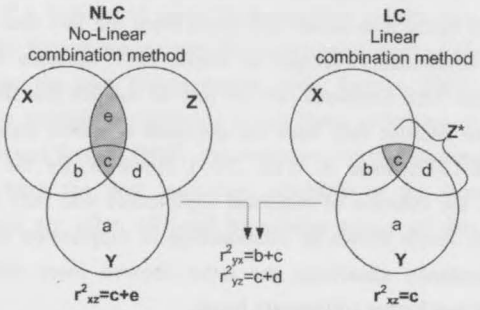



Figure 5.7 Ballantine/Venn diagram giving a tentative illustration of the calculations performed with the Linear combination method (LC) and with the No-Linear combination method (NLC), for the trivariate linear case. Correlations are considered as proportions of variance and are signalled by grey areas. The crude association of X and Z with Y ($r_{yx}$ and $r_{yz}$) do not differ between the two methods, but the collinearity ($r_{xz}$) does. See Cohen et al (2003) for more details on the use of Ballantine diagrams and the "area C problem".

## 5.5 Discussion

The aim of this study was to investigate how frequently different types of confounding arise and what was their impact on statistical significance and on causal interpretation using real-world data; which is inter-connected and where correlations between variables are bounded by physical, biological and socio-economic rules and contingencies. The study is meant only as an exercise for what might be expected in more rigorous and complete environmental epidemiological studies, thus all association estimates should be interpreted with caution.

This study found that suppression situations were not uncommon. Signs of suppression in similar epidemiological studies are difficult to find because researchers rarely report crude associations nowadays, and most use non-linear models where identification of confounding types may differ from the present linear case, although this is a matter in need of more research (e.g.: Lynn, 2003). The only recent report that hinted at this issue was one of the follow-ups of the Washington's US Veterans cohort, which correlated regulated air pollutants and survival (Lipfert et al, 2000). In Table 5 of that article, peak $O_3$ was not a significant predictor of survival ($p>.05$) in simple regression, whereas in multivariate analysis, it became one of the most important predictors, among the pollutants. Both the crude and adjusted effect of peak-$O_3$ were positive, which suggests that the relationships might have been under a PRS situation; however the adjusted effect was estimated with a proportional hazards model and so identification of confounding types, as described in the present work, need not apply.

Out of the 67 relationships analysed, 44 were found to be in a redundancy situation, and 23 in a suppression situation. Suppression in most relationships (17) was not so much due to excessive collinearity as to the very low crude association of the pollutants. In this scenario, only two suppression situations are likely to occur: PRS or NRS (Friedman & Wall, 2005). Unless substantive reasoning can defend the possibility that the confounders completely mask the pollutant's effect on hospitalisations, and thereby lead to low crude associations, these relationships are most likely non-causal and result from the fact that the pollutant explain error variance in the confounders (Tzelgov & Henik, 1991; Maassen & Bakker, 2001). The importance of PRS and NRS situations lies not just on the fact that they challenge a causal interpretation, but also because they have the potential to inflate statistical significance by inflating model fit ($R^2$) (Friedman & Wall, 2005). However, for the 23 relationships in a suppression situation, the inflation of statistical significance was very small, possibly owing to the low collinearity levels shown by relationships in suppression situations compared to relationships in redundancy situations; the latter showed more substantial deflation of statistical significance and higher collinearity levels.

Unless identification of confounding types is performed, one may be misled to report their large adjusted associations, with potentially inflated statistical significance, without further ado. By identifying relationships under suppression, the researcher is better equipped to evaluate whether they are consistent with a causal mechanism, for instance whether it would be acceptable that the pollutant's effect was completely masked by confounding. In addition, the researcher is also equipped to evaluate the extent to which the statistical significance of such relationships might be inflated.

Probably the main aspect that can influence the finding of different types of confounding is the method used to select important relationships, and in this respect this paper followed a common methodology of the epidemiological literature, i.e. select relationships where the pollutant and model are statistically significant, after controlling for a set of confounders selected a priori (e.g.: Lipfert et al, 2000; Jorgensen et al, 2007). One way to prevent suppression situations from occurring, in particular those that involve complete masking of the exposure of interest, is to select relationships on the basis that their crude effect is

numerically large (e.g.: $r_{yx} > |.15|$) and/or statistically significant at some probability level, in addition to requiring that their adjusted effect is statistically significant, as is usual. Even then, suppression situations may still emerge, but they are less likely to be cases of complete masking.

It remains unclear to us, whether adjusted associations should be quantified on the basis of the collinearity between the exposure of interest and all individual confounders (NLC method) or on the basis of the collinearity between the exposure of interest and the linear combination of confounders (LC method). The two methods can, at times, identify different types of confounding for the same relationship. This issue has implications for the Change in Estimate criterion, which appears to be currently based on the NLC method.

### 5.5.1 Strengths and limitations

This study is unlike many epidemiological studies of air pollution for three main reasons: it uses a pure ecological design, linear models, and atmospheric biomonitoring data, which is an indirect indicator of atmospheric pollution, and finally it uses prevalence health data, and no lifestyle confounders. Nevertheless, it is believed that these aspects do not limit the study's results scope and applicability. An aggregate ecological study is nearly equivalent in methods to the second stage of a multi-level study of prospective cohorts (e.g.: Willis et al, 2003), although admittedly the former may require the inclusion of more confounders and thus show greater collinearity than the latter (Morgenstern, 2008; Salway, 2003). Linear estimates, on the other hand, can be accurately converted to log-linear estimates, in most practical cases (e.g. page 89 Cameron and Trivedi 1998). The exposure, health outcome and confounder data, despite their faults were the best indicators available at the time of the study, and epidemiological studies are often afflicted by similar issues of data representativity and quality.

## 5.6 Conclusions

To the best of our knowledge, this is the first study that applied modern techniques for identifying confounding types, to a real dataset in the field of epidemiology.

On the basis of the data used in this study, and in the context of an ecological design and linear modelling, it seems likely that a non-negligible fraction of the effect estimates reported in environmental epidemiological studies may be cases of suppression. Consequently, those effect estimates may be difficult to reconcile with a causal mechanism and their statistical significance may be overstated. It is therefore recommended that identification of not just the extent but the type of confounding affecting epidemiological comparisons becomes routine.

# 6 General Discussion

*For want of a nail the shoe was lost.*
*For want of a shoe the horse was lost.*
*For want of a horse the rider was lost.*
*For want of a rider the battle was lost.*
*For want of a battle the kingdom was lost.*
*And all for the want of a horseshoe nail.*
*English proverb*

This thesis is set in the context of epidemiological studies of air pollution. Recent concerns over the health risks of air pollution have shifted the focus onto its chemical composition, especially metals, due to their inherent toxicity and because they allow discrimination between different pollution sources.

Atmospheric biomonitoring allied to nuclear analytical techniques could be instrumental for epidemiology of air pollution with its wealth of historical data, as well as in future sampling surveys since it enables the measurement of a wide range of chemical elements at sampling densities and geographical scales that are difficult to surpass in terms of cost and man-power.

Although it is tempting to "simply" provide effect estimates for health effects, this thesis has focused instead on a critical approach to some methodological aspects, including: 1) data quality in terms of outliers and noise, 2) representativeness of sampling survey's with respect to the units of analysis, 2) estimation and model selection uncertainty and 3) confounding.

## 6.1 Overview

Chapter 2 uses daily data on regulated air pollutants ($PM_{10}$, $SO_2$, NO, $NO_2$ and $O_3$) and hospital admissions over 5.5 years in Lisbon. In such time-series studies it is believed that short exposures to air pollutants can trigger acute health responses within days. Time-series studies are ecological studies that are very suitable in two ways: first, they can access quite disaggregated data (daily) and second, they compare individuals under a relatively constant background and thus are believed to be less affected by confounding. It is the first issue that concerned the investigations.

The traditional unit of analysis is the day, with lags to account for multiple induction periods in a heterogeneous population. It has often been found that aggregation of the exposure variable over increasingly longer periods (several days or weeks), through moving averages or distributed-lag models, invariably leads to larger slopes than shorter periods. The prevailing explanation for this phenomenon is conceptual: longer-exposure windows are able to capture single-day responses due to single-day exposures at multiple lag-intervals. However there are other more tangible explanations for this phenomenon. Aggregation smoothes the data, discarding errors and noise, and consequently decreases the dispersion of the exposure. Under a surrogacy assumption (i.e. that the aggregated variable contains no

more information with respect to the health outcome than the original variable), a decrease in the variance of the exposure must necessarily increase slopes (Lipfert & Wyzga, 1999).

Conversely, aggregation of the dependent variable, which is rarely performed in the literature, under the same assumption would tend to decrease slopes. A new method where both the exposure and response variable are aggregated with moving averages, was compared with the more conventional method of aggregating only the exposure variable. It was found that the new method leads to regression coefficients that are nearly identical to those of the conventional method, but with greater precision and robustness to data changes. The most likely conclusion, multiple induction periods notwithstanding, is that errors and noise are lost to aggregation, but not the signal. Smoothing data through moving averages with relatively small time windows, are shown to have several advantages. First, the attenuation of extreme values is greater when the latter are isolated than when they are clustered. This differential smoothing is thus sensitive as to whether extreme values are likely to be an episodic error or a real phenomenon. Second, it was shown to lead to more robust estimates than robust regression methods (Tukey and Huber weights). Fourth, when the moving window is of 7 days it provides an easy way to control for day of the week effects.

This is an example of a situation where even though there are strong reasons to believe that associations arise at some disaggregated level (individuals or days), aggregation is beneficial in that it provides identical effect estimates, but with greater precision and robustness. Aggregation does, however, completely hinder our ability to locate or attribute associations to the aggregated units, a problem that is part of the ecological fallacy.

Chapter 3 used solely simulated datasets to explore one of the most fundamental questions: how many samples to take in order to ensure a good representation of a survey and how the answer to this question depends on the size and other characteristics of the sampling unit. This is mostly dedicated to geographical sampling. For epidemiological analytical purposes, it is necessary to accurately determine the average exposure in each sampling unit as they collectively reflect the between-area variance (or survey variance) available for epidemiological comparisons. But it is increasingly realised (Wolterbeek et al, 2010; Salway, 2003) that it is also necessary to accurately determine the within-area variance (or local variance) as this gives a measure of the uncertainty inherent to the survey. The ratio between the two variances provides a measure of the quality of the survey and provides a measure of the extent to which ecological estimates are biased by the aggregation (e.g.: Webster, 2007; Salway, 2003).

This thesis provides numerical recommendations for the sample size required to estimate the survey's variance and the local variances for a range of distributions, margins of error and statistical significance. It also provides a tool (sampling without replacement) to calculate sample sizes that, unlike sample size formulas, can be used irrespective of the distribution (normal, lognormal or others) of the population to estimate the mean, variance and higher moments. The tool does, however, require some sort of sample data to start with.

From these investigations it is clear that the sample sizes used to characterise air pollution exposures in geographical epidemiological studies, be it with instrumental monitoring or biomonitoring, are likely insufficient. Given the large areas often considered in epidemiological studies, the within-area variance in exposure is likely to be substantial, and as a result epidemiological effect estimates are likely to be heavily biased upwards from this source alone (Salway, 2003). Quantification of the within-area variances might provide an estimate of the extent of the bias, but it requires a very large number of samples, whose analytical processing cannot be reduced by composite sampling.

Chapter 4 used municipality-level data on hospital admissions due to circulatory diseases summed over 11 years and the concentration of chemical elements determined through lichen biomonitoring. Single-pollutant linear regression with all confounders selected a priori were used for estimation of associations, variable selection and uncertainty estimation. The parameters of greatest interest are the pollutant's slope and its error.

In such epidemiological investigations three issues are important: 1) the robustness of estimates to sampling variability; 2) robustness of confounder selection to sampling variability and 3) model selection uncertainty. The non-parametric bootstrap was used to investigate these issues by inducing some limited type of sampling variability. For the relationships and data considered, the standard error of the slope was fairly robust, indicating that outliers are unlikely to be influencing the slope. One of the most ink or toner-spending issues in epidemiology is that of variable selection. Like others before, this work found that variable selection in a way that does not induce residual confounding, is unlikely to yield enhanced precision or goodness-of-fit. Thus it is best to select a model a priori on substantive grounds. This was made all the more clear considering the multiple testing involved in variable selection (i.e. model selection uncertainty), which tends to render the models' p-values meaningless. This is only one of the fields where there is clear dissonance between the aim of epidemiology and the methods it uses to achieve those aims. Statistical analyses are, for the most part, unsuited and even antagonist to causal analyses, even though deep down statistical analyses are concerned with causal associations, the latter cannot be determined from observational data alone. However, the method used to assess model selection uncertainty does not appear to be entirely adequate for the purpose. Confounder selection was remarkably vulnerable to sampling variability but was consistent within hospitalisation categories. The latter seems thus more suited for model averaging.

Chapter 5 uses the same basic relationships and analyses as the fourth chapter to identify the confounding situations afflicting each relationship. Confounding control is often conveyed as an intensely conservative procedure whereby the effect (as measured by the slope) is reduced and ultimately prevails. However, there are situations where the opposite can occur, this is known as negative confounding or suppression. There are several reasons to distinguish between different types of confounding, especially in the context of ecological studies: 1) negative confounding decreases errors and increases goodness-of-fit increasing the chance of spurious associations, whereas positive confounding has the opposite effect and 2) negative confounding is a likely scenario when the exposure of interest is a weak predictor of the

dependent variable (in terms of beta-weight), a situation that is very common with air pollutants. Given the wide variety of elements under analysis, it would be expected that they would correlate with confounders in a way that would lead to both negative or positive confounding. Among the studied relationships most were under redundancy, followed by positive reciprocal suppression and negative reciprocal suppression. In many relationships under suppression the crude association of the pollutant with the health outcome was nearly zero and non-significant, whereas its adjusted effect was large and significant. This may be interpreted in two ways: either the pollutant was completely masked by the cofounders or the pollutant was not associated with the health outcome but instead explains error variance in the confounders. It is unclear in many epidemiological studies whether air pollutants are selected on the basis of their crude association with the health outcome, prior to multivariate analyses or not (exception Lipfert et al, 2003).

## 6.2 Final Remarks

It is common wisdom that often the most important fundamental things are the most undervalued and overlooked. This is clearly the case with data collection (the nail in the battle). Sampling and analyses are the raw materials of epidemiological analyses, and no statistical procedure can replace good quality data.

Several authors have hinted, in different contexts, at the idea that the data routinely used in environmental epidemiological studies (even multi-level ones) is not worthy of the arsenal of increasingly sophisticated analytical techniques that are inflicted upon it (e.g.: Wakefield, 2003, 2008; Hayes, 2003; Chatfield, 1995). The increased computational power, the development of statistical algorithms and their prompt inclusion in commercial statistical software have greatly expanded the panoply of analytical techniques available to researchers. The multiple testing inherent to trying different approaches to the data means that p-values begin to be meaningless. P-values are already widely disregarded in epidemiology, but what is there to replace it? Equivalent progresses in data quantity, quality and understanding has not been observed.

It is somewhat perplexing to realise that epidemiological comparisons are often made across units which are more variable within themselves than between themselves. What is the meaning of such comparisons (Salway, 2003)?

Regulated air pollutants, and possibly also airborne chemical elements, are minor health risk factors compared to other factors, both lifestyle and socioeconomic. If disease reduction is the goal of public health, environmental air pollution does not strike as a priority, nowadays in the developed world. However, as regulations tighten and industry and emission sources struggle to keep up with them and develop new manufacturing techniques and new fields, other unregulated and/or unknown air pollutants may be emitted in increasingly larger amounts (e.g.: ultrafine particles) at the cost of abating the air pollutants that are under regulation. There is therefore, the need for some degree of suspicion as to what might be contaminating our environment. To keep track and monitor such a wide range of chemicals is clearly an impossibility; however with regards to several pollutants, including chemical

elements, environmental biomonitoring is a canny if unpolished supplementary approach to instrumental monitoring.

## 6.3 Future Research

The basic recommendations that can be derived from this study necessarily pertain to directing efforts towards data of greater quality.

Atmospheric biomonitoring surveys should be carried out with epidemiological purposes in mind and large concerted efforts should be made to harmonise and ensure good sampling and analytical procedures, perhaps even by having all analyses performed in a single laboratory (Wolterbeek et al, 2010). Ensuring that local variances are much smaller than the survey's variance is one important aim of future surveys. Thus future sampling surveys will necessarily have to be much denser, tailored to areas where populations concentrate and thus health events are more numerous, and will necessarily have to be complemented by instrumental monitoring to at least calibrate biomonitor's accumulation to environmental factors and time.

Health institutes would, in an ideal world, would be able to provide health data at any desired geographical scale and division required by the epidemiologist. The best way to minimise ecological bias is to create the areas of comparison in a way that maximises the between-area variance and minimises the within-area variance in exposure. This also reduces (potentially) the number of samples and of analyses. Health data should be able to match such defined geographical areas.

Diagnosis of confounding situations should be made for all relationships because large collinearity and/or weak pollutant's effects can give rise to paradoxical situations which appear to be difficult to reconcile with causal mechanisms and that could lead to too optimistic errors and model fit.

# List of abbreviations

## Chemical elements

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Al | Aluminium | Cs | Caesium | Mg | Magnesium | Sm | Samarium |
| As | Arsenic | Cu | Copper | Mn | Manganese | Sn | Tin |
| Au | Gold | Eu | Europium | Mo | Molybdenium | Sr | Strontium |
| Ba | Barium | Fe | Iron | Na | Sodium | Tb | Tiberium |
| Be | Berilium | Ga | Gallium | Nd | Neodymium | Th | Thorium |
| Br | Bromine | Ge | Germanium | Ni | Nickel | Ti | Titanium |
| Ca | Calcium | Hf | Hafnium | Pb | Lead | Tl | Thallium |
| Cd | Cadmium | Hg | Mercury | Rb | Rubidium | U | Uranium |
| Ce | Cerium | I | Iodine | S | Sulphur | V | Vanadium |
| Cl | Chlorine | K | Potassium | Sb | Antimony | W | Tungsten |
| Co | Cobalt | La | Lanthanum | Sc | Scandium | Yb | Ytterbium |
| Cr | Chromium | Lu | Lutetium | Se | Selenium | Zn | Zinc |

## Chapter 1

AIC – Akaike Information Criterion
ATSDR – Agency for Toxic Substances & Disease Registry (USA)
BIC – Bayesian Information Criterion
BS – British Smoke
CE – Change in Estimate criterion
CO – Carbon Monoxide
EPA IRIS – Environmental Protection Agency Integrated Risk Information (USA)
IAEA – International Atomic Energy Agency
IARC – International Agency for Research on Cancer
MCTTFA - Monte Carlo Target Transform Factor Analysis
PAH – Polycyclic Aromatic Hydrocarbons (e.g.: benzene)
PM – general particulate matter of <10μm aerodynamic size, includes PM10 and PM2.5
PM10 – Particulate matter of <10μm aerodynamic size
PM2.5 - Particulate matter of <2.5μm aerodynamic size
NO – Nitrogen Oxide
NO2 – Nitrogen Dioxide
NUTS-III – French acronym for Nomenclature of Territorial Units for Statistics, level 3.
O3 – Ozone
SO2 – Sulphur Dioxide
TSP – Total Suspended Particles
UN – United Nations
WHO – World Health Organisation

## Chapter 2

ACSS – Administração Central do Sistema de Saúde
AP – Air pollutant (PM10, SO2, NO, NO2, CO and O3)
Bint- slope of a regression between the (mean) intercept given by one model and the (mean) intercept given by an alternative model, over a number of relationships
Bslopes- slope of a regression between the slopes given by one model and the slopes given by an alternative model, over a number of relationships
CO-Carbon Monoxide
CMA – Centred 7-day moving average
CMA & CMA – model where both the health variable and the air pollutant are expressed as 7-day centred moving averages
DLM – Distributed lag model

FMA – Forward 7-day moving average
FMA&PMA - model where the health variable is expressed with a 7-day forward moving average and the air pollutant is expressed with a 7-day prior moving average
HA – Hospital admissions
ICD9-CM – International Classification of Diseases, 9th revision, Clinical Modification.
MA – Moving averages
M-estimation – robust estimation (iterative) based on weights such as Tukey's and Huber's
MLE – Maximum Likelihood Estimate
NO-Nitrogen Oxide
NO2-Nitrogen Dioxide
O – original daily value (no moving averages)
O3-Ozone
O&CMA – model where the health variable is expressed with its daily value whereas the air pollutant is expressed as 7-day centred moving average
O&DLM - model where the health variable is expressed with its daily value whereas the air pollutant is expressed as a distributed lag model of 7 days
O&PMA – model where the health variable is expressed with its daily value whereas the air pollutant is expressed with a 7-day prior moving average
OLS – Ordinary Least Squares
PM10 – Particulate matter with aerodynamic size less than 10μm
PM2.5 – Particulate matter with aerodynamic size less than 2.5μm
PMA – Prior 7-day moving average
RF – risk factors
SO2- Sulphur Dioxide
WHO – World Health Organisation

## Chapter 3

B – Random-within-blocks sampling
B/W – between variance to within variance ratio
G – Systematic-grid sampling
LV - local variance
MSS – minimum sample size (usually refererring to the purpose of estimating the mean or the variance.
ME – margin of error
N – number o observations
P10, P25, P75, P90 – percentiles
R – Simple random sampling
RSD – relative standard deviation in %
SV/LV – survey variance to local variance ratio
WA – within-area variance in %

## Chapter 4

ACSS – Administração Central do Sistema de Saúde
aRD – Risk Difference per achievable change in chemical element
aRR – Risk Rato per achievable change in chemical element
B – Backward model or naïve (unstandardised) slope of chemical element, should be clear from context
B* - bootstrap (unstandardised) slope of chemical element
B/W - between-area to within-area variance ratio (also known as signal-to-noise ratio)
C1 to C33 – confounders
CBV – cerebrovascular diseases (ICD9-CM: 430-438)
CE – change in estimate criterion
CIRC – circulatory diseases (ICD9-CM: 390-459)
E – Elasticity at the mean
F – Full model or Females, should be clear from context
F1, F2 and F5 - selected emission factors calculated from chemical elements database by MCTTFA.
ICD9-CM – International Classification of Diseases 9th revision, Clinical Modification
IHD – ischemic heart diseases (ICD9-CM: 410-414)
$k_0$-INAA – Instrumental Neutron Activation Analysis with $k_0$ method
M - males

MB – Mean Backward model
MCTTFA – Monte Carlo Target Transform Factor Analysis
OLS – Ordinary least squares
PIXE – Particle-Induced X-ray Emission
$R^2$ – multiple correlation coefficient
S – Simple model
SE*(E) – bootstrap standard error of the elasticity at the meant
SE – naïve standard error of the slope of chemical element
SE* - bootstrap standard error of the slope of chemical element
SE(E) – naïve standard error of the elasticity at the mean

## Chapter 5

CBV – Cerebrovascular diseases (ICD9-CM: 430-438)
CIRC – Circulatory diseases (ICD9-CM: 390-459)
F1, F2, F3 and F5 – selected emission factors calculated from chemical elements database by MCTTFA.
ICD9-CM – International Classification of Diseases 9th revision, Clinical Modification
IHD – Ischemic heart disease (ICD9-CM: 410-414)
LC – linear combination method
NLC – no linear combination method
NRS – negative reciprocal suppression
NS – negative suppression
PRS – positive reciprocal suppression
$R^2$ – multiple correlation coefficient
SE – standard error of beta-weight
Y – dependent variable
X – exposure of interest
Z – confounder
Z* - linear combination of 30 or 33 confounders

# References


Abbey DE, Nishino N, McDonnell WF, Burchette RJ, Knutsen SF, Lawrence WL, Yang JX. **1999**. Long-Term inhalable particles and other air pollutants related to mortality in nonsmokers. Am J Respir Crit Care Med 159:373-382.

Aboal JR, Real C, Fernandez JA, Carballeira A. **2006**. Mapping the results of extensive surveys: The case of atmospheric biomonitoring and terrestrial mosses. Sci Total Environ 356:256-274.

ATSDR, Agency for Toxic Substances and Disease Registry US Department of Health and Human Services. **2007**. *Toxicological Profile for Arsenic*. 1-559.

ATSDR. Last accessed: **Nov. 2011**. Toxic Substances Portal. Available at: http://www.atsdr.cdc.gov/toxprofiles/index.asp.

Avakian MD, Dellinger B, Fiedler H, Gullet B, Koshland C, Marklund S, Oberdorster G, Safe S, Sarofim AF, Smith KR, Schwartz D, Suk WA. **2002**. The origin, fate, and health effects of combustion by-products: a research framework. Environ Health Perspect 110:1155-1162.

Bates DV. **1992**. Health indices of the adverse effects of air pollution: the question of coherence. Environ Res 59:336-349.

Baxter LA, Finch SJ, Lipfert FW, Yu Q. **1997**. Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. Risk Anal 17:273-278.

Best N, Cockings S, Bennett J, Wakefield J, Elliott P. **2001**. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. J R Stat Soc Ser A 164:155-174.

Blot WJ, Fraumeni JFJr. **1977**. Geographic patterns of oral cancer in the United States: Etiologic implications. Journal of Chronic Diseases 30:745-757.

Bluestone B, Harrison B. **1982**. The deindustrialisation of America: plant closings, community abandonment, and the dismantling of basic industry. New York: Basic Books, Inc.

Braga ALF, Saldiva PHN, Pereira LAA, Menezes JJC, Conceição GMS, Lin CA, Zanobetti A, Schwartz J, Dockery DW. **2001**. Health effects of air pollution exposure on children and adolescents in São Paulo, Brazil. Pediatr Pulmonol 31:106-113.

Branquinho C, Catarino F, Brown DH, Pereira MJ, Soares A. **1999**. Improving the use of lichens as biomonitors of atmospheric metal pollution. Sci Total Environ 232:67-77.

Breiman L. **1992**. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. J Am Stat Assoc 87:738-754.

Brenner H, Savitz DA, Jockel KH, Greenland S. **1992**. Effects of nondifferential exposure misclassification in ecologic studies. Am J Epidem 135:85-95.

Brunekreef B, Holgate ST. **2002**. Air pollution and health. Lancet 360:1233-1242.

Buse A, Norris D, Harmens H, Buker P, Ashenden TW, Mills G. **2003**. *Heavy metals in European mosses 2000/2001 survey*. UNECE ICP Vegetation 1-46.

Cameron AC, Trivedi PK. **1998**. Regression analysis of count data. Cambridge University Press.

Chatfield C. **1995**. Model uncertainty, data mining and statistical inference. J R Stat Soc Ser A 158:419-466.

Chaudhuri A, Stenger H. **2005**. Survey sampling: theory and methods. Chapman & Hall/CRC.

Chen C, Chock DP, Winkler SL. **1999**. A simulation study of confounding in Generalized Linear Models for air pollution epidemiology. Environ Health Perspect 107:217-222.

Chernick MR. **2008**. Bootstrap methods: a guide for practitioners and researchers. John Wiley & Sons, Inc.

Cislaghi C, Nimis PL. **1997**. Lichens, air pollution and lung cancer. Nature 387:463-464.

Claiborn CS, Larson TV, Sheppard L. **2002**. Testing the metals hypothesis in Spokane, Washington. Environ Health Perspect 110:547-552.

Cochran WG. **1977**. Sampling techniques. John Wiley & Sons, Inc.

Cohen BL. **1994**. Invited Commentary: In defense of ecologic studies for testing a linear-no threshold theory. Am J Epidem 139:765-768.

Cohen J, Cohen P, West SG, Aiken LS. **2003**. Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum Associates, Inc.

Conti ME, Cecchetti G. **2001**. Biological monitoring: lichens as bioindicators of air pollution assessment: a review. Environ Pollut 114:471-492.

Costa DL. **1998**. Ambient particulate matter: Is there a toxic role for constitutive transition metals? Studies in Environmental Science 72:117-123.

Costa DL, Dreher KL. **1999**. What do we need to know about airborne particles to make effective risk management decisions? A toxicology perspective. Hum Ecol Risk Assess 5:481-491.

Cox LH. **2000**. Statistical issues in the study of air pollution involving airborne particulate matter. Environmetrics 11:611-626.

Cressie NAC. **1993**. Statistics for spatial data. John Wiley & Sons, Inc.

Dockery DW, Pope CAI, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BGJr, Speizer FE. **1993**. An association between air pollution and mortality in six U.S. cities. N Engl J Med 329:1753-1759.

Dreher KL. **2000**. Particulate matter physicochemistry and toxicology: In search of causality - A critical perspective. Inhal Toxicol 12:45-57.

Durkheim E. **1979**. Suicide: A study in Sociology. Macmillan Publishing Co., Inc.

Dusseldorp A, Kruize H, Brunekreef B, Hofschreuder P, de Meer G, van Oudvorst AB. **1995**. Associations of PM10 and airborne iron with respiratory health of adults living near a steel factory. Am J Respir Crit Care Med 152:1932-1939.

Dye JA, Adler KB, Richards JH, Dreher KL. **1997**. Epithelial injury induced by exposure to residual oil fly-ash particles: Role of reactive oxygen species? Am J Respir Cell Mol Biol 17:625-633.

Dye JA, Adler KB, Richards JH, Dreher KL. **1999**. Role of soluble metals in oil fly ash-induced airway epithelial injury and cytokine gene expression. Am J Physiol 277:L498-L510.

Dye JA, Lehmann JR, McGee JK, Winsett DW, Ledbetter AD, Everitt JI, Ghio AJ, Costa DL. **2001**. Acute pulmonary toxicity of particulate matter filter extracts in rats: coherence with epidemiologic studies in Utah Valley residents. Environ Health Perspect 109:395-403.

EEA. **1998**. *Guidance report on preliminary assessment under EC air quality directives*. Copenhagen: European Environmental Agency Technical Report N11,

Efron B, Tibshirani RJ. **1993**. An introduction to the bootstrap. Chapman & Hall/CRC.

EIONET. Last accessed: **Nov. 2011**. AirBase - the European Air quality dataBase. Available at: http://acm.eionet.europa.eu/databases/airbase/.

El Khoukhi T, Cherkaoui RM, Gaudry A, Ayrault S, Senhou A, Chouak A, Moutia Z, Chakir E. **2004**. Air pollution biomonitoring survey in Morocco using $k_0$-INAA. Nucl Instrum Meth B 213:770-774.

Ellison G, Newham J, Pinchin MJ, Thompson I. **1976**. Heavy metal content of moss in the region of Consett (North East England). Environ Pollut 11:167-174.

Faraway JL. **1992**. On the cost of data analysis. J Comput Graph Stat 1:213-229.

Fewell Z, Davey SG, Sterne JAC. **2007**. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. Am J Epidem 166:646-655.

Figueira R, Sergio C, Sousa AJ. **2002**. Distribution of trace metals in moss biomonitors and assessment of contamination sources in Portugal. Environ Pollut 118:153-163.

Firebaugh G. **1978**. A rule for inferring individual-level relationships from aggregate data. Am Sociol Rev 43:557-572.

Freitas MC, Reis MA, Alves LC, Wolterbeek HTh, Verburg TG, Gouveia M. **1997**. Biomonitoring of trace-element air pollution in Portugal: Qualitative survey. J Radioanal Nucl Ch 217:21-30.

Freitas MC, Reis MA, Alves LC, Wolterbeek HTh. **1999**. Distribution in Portugal of some pollutants in the lichen *Parmelia sulcata*. Environ Pollut 106:229-235.

Freitas MC, Reis MA, Alves LC, Wolterbeek HTh. **2000**. Nuclear analytical techniques in atmospheric trace element studies in Portugal. In: Markert BA, Friese K, eds. Trace Elements: Their Distribution and Effects in the Environment. Elsevier, 187-213.

Friedman L, Wall M. **2005**. Graphical views of suppression and multicollinearity in multiple linear regression. Am Stat 59:127-136.

Gailey FAY, Lloyd OL. **1993**. Spatial and temporal patterns of airborne metal pollution: the value of low technology sampling to an environmental epidemiology study. Sci Total Environ 133:201-219.

Garty J. **2001**. Biomonitoring atmospheric heavy metals with lichens: Theory and application. Crit Rev Plant Sci 20:309-371.

Garty J, Lehr H, Garty-Spitz RL. **2009**. Three decades of biomonitoring airborne Pb in a rural area with the epiphytic lichen *Ramalina lacera*: A retrospective study. Israel J Plant Sci 57:25-34.

Gavett SH, Haykal-Coates N, Copeland LB, Heinrich J, Gilmour MI. **2003**. Metal composition of ambient PM2.5 influences severity of allergic airways disease in mice. Environ Health Perspect 111:1471-1477.

Gerlofs-Nijland ME, Rummelhard M, Boere AJ, Leseman DLAC, Duffin R, Schins RPF, Borm PJA, Sillanpaa M, Salonen RO, Cassee FR. **2009**. Particle induced toxicity in relation to transition metal and polycyclic aromatic hydrocarbon contents. Environ Sci Technol 43:4729-4736.

Ghio AJ, Stonehuerner J, Dailey LA, Carter JD. **1999**. Metals associated with both the water-soluble and insoluble fractions of an ambient air pollution particle catalyze an oxidative stress. Inhal Toxicol 11:37-49.

Ghio AJ, Devlin RB. **2001**. Inflammatory lung injury after bronchial instillation of air pollution particles. Am J Respir Crit Care Med 164:704-708.

Ghio AJ, Silbajoris R, Carson JL, Samet JM. **2002**. Biologic effects of oil fly ash. Environ Health Perspect 110:89-94.

Ghio AJ. **2004**. Biological effects of Utah Valley ambient air particles in humans: a review. J Aerosol Med 17:157-164.

Glynn A, Wakefield J, Handcock MS, Richardson S. **2008**. Alleviating linear ecological bias and optimal design with sub-sample data. J R Stat Soc Ser A 171:179-202.

Godinho RM, Freitas MC, Wolterbeek HTh. **2004**. Assessment of lichen vitality during a transplantation experiment to a polluted site. J Atmos Chem 49:355-361.

Godinho RM, Wolterbeek HT, Verburg T, Freitas MC. **2008**. Bioaccumulation behaviour of transplants of the lichen *Flavoparmelia caperata* in relation to total deposition at a polluted location in Portugal. Environ Pollut 151:318-325.

Godinho RM, Verburg TG, Freitas MC, Wolterbeek HTh. **2009a**. Accumulation of trace elements in the peripheral and central parts of two species of epiphytic lichens transplanted to a polluted site in Portugal. Environ Pollut 157:102-109.

Godinho RM, Wolterbeek HTh, Pinheiro M, Alves LC, Verburg TG, Freitas MC. **2009b**. Micro-scale elemental distribution in the thallus of *Flavoparmelia caperata* transplanted to polluted site. J Radioanal Nucl Ch 281:205-210.

Godinho RM. **2010**. *Lichen biomonitors: factors affecting response behaviour.* PhD, Delft University of Technology.

Godinho RM, Verburg TG, Freitas MC, Wolterbeek HTh. **2011a**. Assessment of acid-base buffering properties of *Flavoparmelia caperata*: influence of age and pollution exposure. International Journal of Environment and Health 5:19-31.

Godinho RM, Verburg TG, Freitas MC, Wolterbeek HTh. **2011b**. Dynamics of element accumulation and release of *Flavoparmelia caperata* during a long-term field transplant experiment. International Journal of Environment and Health 5:49-59.

Goldberger AS. **1991**. A course in Econometrics. Harvard University Press.

Goodman PG, Dockery DW, Clancy L. **2004**. Cause-specific mortality amd the extended effects of particulate pollution and temperature exposure. Environ Health Perspect 112:179-185.

Greenbaum D. **2003**. A historical perspective on the regulation of particles. J Toxicol Environ Health A 66:1493-1498.

Greenland S. **1980**. The effects of misclassification in the presence of covariates. Am J Epidem 112:564-569.

Greenland S, Morgenstern H. **1989**. Ecological bias, confounding, and effect modification. Int J Epidemiol 18:269-274.

Greenland S. **1989**. Modeling and variable selection in epidemiologic analysis. Am J Public Health 79:340-349.

Greenland S, Brenner H. **1993**. Correcting for non-differential misclassification in ecologic analyses. J R Stat Soc Ser C Appl Stat 42:117-126.

Greenland S, Robins CJ. **1994a**. Ecologic studies, biases, misconceptions and counterexamples. Am J Epidem 139:747-760.

Greenland S, Robins JM. **1994b**. Accepting the limits of ecologic studies: Drs. Greenland and Robins reply to Drs. Piantadosi and Cohen. Am J Epidem 139:769-771.

Greenland S, Morgenstern H. **2001**. Confounding in health research. Annu Rev Public Health 22:189-212.

Hale WE. **1972**. Sample size determination for the log-normal distribution. Atmos Environ 6:419-422.

Harmens H, Buse A, Buker P, Norris D, Mills G, Williams B, Reynolds B, Ashenden TW, Ruhling A, Steinnes E. **2004**. Heavy Metal Concentrations in European Mosses: 2000/2001 Survey. J Atmos Chem 49:425-436.

Harmens H. **2010**. *Heavy Metals in Europena Mosses: 2010 Survey. Monitoring Manual. Monitoring of atmospheric deposition of heavy metals, nitrogen adn POPs in Europe using Bryophytes.* UNECE ICP Vegetation 1-16.

Harrison RM, Yin J. **2000**. Particulate matter in the atmosphere: which particle properties are important for its effects on health? Sci Total Environ 249:85-101.

Harrison RM, Smith DJT, Kibble AJ. **2004**. What is responsible for the carcinogenicity of PM2.5? Occup Environ Med 61:799-805.

Hayes M. **2003**. "Ecologic Confounders" in the context of a spatial analysis of the air pollution-mortality relationship. J Toxicol Environ Health A 66:1779-1782.

HEI. **2002**. *Understanding the health effects of components of the particulate matter mix: progress and next steps*. HEI 2, 1-20.

Heller-Zeisler SF, Zeisler R, Zeiller E, Parr RM, Radecki Z, Burns I, Regge PD. **1999**. *Report on the intercomparison run for the determination of trace and minor elements in lichen material IAEA-336*. Austria: IAEA IAEA/AL/079 NAHRES-33, 1-90.

Heymans M, van Buuren S, Knol D, van Mechelen W, de Vet H. **2007**. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Methodol 7:33-

Holling H. **1983**. Suppressor structures in the General Linear Model. Educ Psychol Meas 43:1-9.

INE. **2006**. *O País em Números. Informação Estatística 1991 a 2004 (CD-ROM)*. Portugal: INE.

INE, INSA. **2009**. Inquérito Nacional de Saúde 2005/2006. Lisboa: INE, INSA.

INE. Last accessed: **Jan. 2011**. Statistical Data. Available at: http://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE.

Ito K, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV, Neas LM, Hopke PK, Thurston GD. **2006**. PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. J Expo Sci Environ Epidemiol 16:300-310.

Jackson C, Best N, Richardson S. **2006**. Improving ecological inference using individual-level data. Stat Med 25:2136-2159.

Jeran Z, Jacimovic R, Batic F, Smodis B, Wolterbeek HTh. **1996**. Atmospheric heavy metal pollution in Slovenia derived from results for epiphytic lichens. Fresenius J Anal Chem 354:681-687.

Jeran Z, Jacimovic R, Mikuz PP. **2003**. Lichens and mosses as biomonitors. J Phys IV 107:675-678.

Jerrett M, Burnett R, Willis A, Krewski D, Goldberg MS, DeLuca P, Finkelstein JN. **2003**. Spatial analysis of the air pollution-mortality relationship in the context of ecologic confounders. J Toxicol Environ Health A 66:1735-1778.

Jorgensen E, Keiding N, Grandjean P, Weihe P. **2007**. Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. Ann Epidemiol 17:27-35.

Kennedy NJ, Hinds WC. **2002**. Inhalability of large solid particles. J Aerosol Sci 33:237-255.

Kennedy T, Ghio AJ, Reed W, Samet JM, Zagorski J, Quay J, Carter JD, Dailey LA, Hoidal JR., Devlin RB. **1998**. Copper-dependent inflammation and nuclear factor-kappaB activation by particulate air pollution. Am J Respir Cell Mol Biol 19:366-378.

Knaapen AM, Shi T, Borm PJA, Schins RPF. **2002**. Soluble metals as well as the insoluble particle fraction are involved in cellular DNA damage induced by particulate matter. Mol Cell Biochem 234-235:317-326.

Kodavanti UP, Hauser R, Christiani DC, Meng ZH, McGee JK, Ledbetter AD, Richards JH, Costa DL. **1998**. Pulmonary responses to oil fly ash particles in the rat differ by virtue of their specific soluble metals. Toxicol Sci 43:204-212.

Krewski D, Burnett R, Goldberg MS, Hoover BK, Siemiatycki J, Jerrett M, Abrahamowicz M, White WH, et al. **2000**. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of particulate air pollution and mortality. Part II: Sensitivity analyses. Appendix E: Selection of ecologic covariates for the ACS study. (available on request). Health Eff Inst, Spec Rep:1-23.

Krewski D, Burnett R, Goldberg MS, Hoover BK, Siemiatycki J, Jerrett M, Abrahamowicz M, White WH. **2003**. Overview of the Reanalysis of the Harvard Six Cities Study and American Cancer Society Study of particulate air pollution and mortality. J Toxicol Environ Health A 66:1507-1552.

Krewski D, Burnett R, Jerrett M, Pope CAI, Rainham D, Calle EE, Thurston G, Thun M. **2005**. Mortality and long-term exposure to ambient air pollution: ongoing analyses based on the American Cancer Society cohort. J Toxicol Environ Health A 68:1093-1109.

Kuik P, Blaauw M, Sloof JE, Wolterbeek HTh. **1993**. The use of Monte Carlo methods in factor analysis. Atmos Environ 27:1967-1974.

Kuik P, Sloof JE, Wolterbeek HTh. **1993**. Application of Monte Carlo-assisted factor analysis to large sets of environmental pollution data. Atmos Environ 27:1975-1983.

Kuik P, Wolterbeek HTh. **1995**. Factor analysis of atmospheric trace-element deposition data in the netherlands obtained by moss monitoring. Water Air Soil Poll 84:323-346.

Laden F, Neas LM, Dockery DW, Schwartz J. **2000**. Association of fine particulate matter from different sources with daily mortality in Six U.S. Cities. Environ Health Perspect 108

Lee CSL, Li X, Zhang G, Peng X, Zhang L. **2005**. Biomonitoring of trace metals in the atmosphere using moss (*Hypnum plumaeforme*) in the Nanling Mountains and the Pearl River Delta, Southern China. Atmos Environ 39:397-407.

Lighty JS, Veranth JM, Sarofim AF. **2000**. Combustion aerosols: factors governing their size and composition and implications to human health. Air Waste 50:1619-1622.

Lipfert FW. **1980**. Statistical studies of mortality and air pollution multiple regression analyses by cause of death. Sci Total Environ 16:165-183.

Lipfert FW. **1988**. *A statistical study of the macroepidemiology of air pollution and total mortality*. New York: Brookhaven National Lab BNL-52122, 1-143.

Lipfert FW. **1993**. A critical review of studies of the association between demands for hospital services and air pollution. Environ Health Perspect 101:229-268.

Lipfert, F. W. **1995**. "Estimating air pollution-mortality risks from cross-sectional studies: prospective vs ecologic study designs", in *In: Particulate matter: health and regulatory issues: proceedings of an international specialty conference.*, A&WMA publication, Pittsburgh, PA., pp. 78-102.

Lipfert FW, Wyzga RE. **1995a**. Air pollution and mortality: issues and uncertainties. Air Waste 45:949-966.

Lipfert FW, Wyzga RE. **1995b**. Uncertainties in identifying responsible pollutants in observational epidemiology studies. Inhal Toxicol 7:671-689.

Lipfert FW. **1997**. Air pollution and human health: Perspectives for the 90s and beyond. Risk Anal 17:137-146.

Lipfert FW. **1998**. Trends in airborne particulate matter in the United States. Appl Occup Environ Hyg 13:370-384.

Lipfert FW. **1999**. The use and misuse of surrogate variables in environmental epidemiology. J Environ Med 1:267-278.

Lipfert FW, Perry HMJr, Miller JP, Baty JD, Wyzga RE, Carmody SE. **2000**. The Washington University EPRI Veteran's cohort mortality study: Preliminary Results. Inhal Toxicol 12:41-73.

Lipfert FW, Morris SC. **2002**. Temporal and spatial relations between age specific mortality and ambient air quality in the United States: regression results for counties, 1960-97. Occup Environ Med 59:156-174.

Lipfert FW, Perry HMJr, Miller JP, Baty JD, Wyzga RE, Carmody SE. **2003**. Air pollution, blood pressure, and their long-term associations with mortality. Inhal Toxicol 215:493-512.

Lipfert FW, Baty JD, Miller JP, Wyzga RE. **2006**. PM2.5 constituents and related air quality variables as predictors of survival in a cohort of US Military Veterans. Inhal Toxicol 18:645-657.

Lumley T, Sheppard L. **2000**. Assessing seasonal confounding and model selection bias in air pollution epidemiology using positive and negative control analyses. Environmetrics 11:705-717.

Lynn HS. **2003**. Suppression and confounding in action. Am Stat 57:58-61.

Maassen GH, Bakker AB. **2001**. Suppressor variables in path models. Sociol Methods Res 30:241-270.

MacKinnon DP, Krull JL, Lockwood CM. **2000**. Equivalence of the mediation, confounding and suppression Effect. Prev Sci 1:173-181.

Maldonado G, Greenland S. **1993**. Simulation study of confounder-selection strategies. Am J Epidem 138:923-936.

Mar TF, Ito K, Koenig JQ, Larson TV, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Neas LM, Stolzel M, Paatero P, Hopke PK, Thurston GD. **2006**. PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of PM2.5 and daily mortality in Phoenix, AZ. J Expo Sci Environ Epidemiol 16:311-320.

Markert BA, Herpin U, Siewers U, Berlekamp J, Lieth H. **1996**. The German heavy metal survey by means of mosses. Sci Total Environ 182:159-168.

Markert BA, Breure AM, Zechmeister HG. **2003**. Definitions, strategies and principles for bioindication/biomonitoring of the environment. In: Markert BA, Breure AM, Zechmeister HG, eds. Bioindicators & Biomonitors: Principles, concepts and applications. Elsevier, 3-39.

Marques APVM, Freitas MC, Reis MA, Wolterbeek HTh, Verburg TG, de Goeij JJM. **2004**. Lichen-Transplant Biomonitoring in the Assessment of Dispersion of Atmospheric Trace-Element Pollutants: Effects of Orientation Towards the Wind Direction. J Atmos Chem 49:211-222.

Marques APVM. **2008**. *Positional responses in lichen transplant biomonitoring of trace element air pollution.* PhD, Delft University of Technology.

Maynard R, Cohen A. **2003**. Public Health Significance of Research Results. J Toxicol Environ Health A 66:1877-1878.

McNamee R. **2003**. Confounding and confounders. Occup Environ Med 60:227-234.

Mickey RM, Greenland S. **1989**. The impact of confounder selection on effect estimation. Am J Epidem 129:125-137.

Morgenstern H. **1982**. Uses of ecologic analysis in epidemiologic research. Am J Public Health 72:1336-1344.

Morgenstern H. **2008**. Ecologic Studies. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology. Philadelphia: Lippincot Williams & Wilkins, 511-531.

NRC. **1998**. Research priorities for airborne particulate matter I: Immediate priorities and a long-range research portfolio. Washington DC: National Academy Press.

Nyarko BJB, Adomako D, Serfor-Armah Y, Dampare SB, Adotey D, Akaho EHK. **2006**. Biomonitoring of atmospheric trace element deposition around an industrial town in Ghana. Radiat Phys Chem 75:954-958.

Nylander W. **1886**. Les lichens du Jardin du Luxembourg. B Soc Bot Fr 13:364-

Oakes JM. **2009**. Commentary: Individual, ecological and multilevel fallacies. Int J Epidemiol 38:361-368.

Oberdorster G, Utell MJ. **2002**. Ultrafine particles in the urban air: to the respiratory tract and beyond? Environ Health Perspect 110:A440-A441.

Pearl J. **1995**. Causal diagrams for empirical research. Biometrika 82:669-688.

Pearl J. **1998**. *Why there is no statistical test for confounding, why many think there is, and why they are almost right.* Technical Report R-256, 1-13.

Pearl J. **2000**. Causality: models, reasoning and inference. Cambridge University Press.

Piantadosi S. **1994**. Invited Commentary: Ecologic Biases. Am J Epidem 139:761-764.

Pignata ML, Pla RR, Jasan RC, Martinez MS, Rodriguez JH, Wannaz ED, Gudino GL, Carreras HA, Gonzalez CM. **2007**. Distribution of atmospheric trace elements and assesment of air quality in Argentina employing the lichen, *Ramalina celastri*, as a passive biomonitor: detection of air pollution emission sources. International Journal of Environment and Health 1:29-46.

Pitard A, Viel JF. **1997**. Some methods to address collinearity among pollutants in epidemiological time series. Stat Med 16:527-544.

Pope CAI. **1989**. Respiratory disease associated with community air pollution and a steel mill, Utah Valley. Am J Public Health 79:623-628.

Pope CAI. **1991**. Respiratory hospital admissions associated with PM10 pollution in Utah, Salt Lake, and Cache Valleys. Arch Environ Health 46:90-97.

Pope CAI, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CWJr. **1995**. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med 151:669-674.

Pope CAI. **1996**. Particulate pollution and health: a review of the Utah valley experience. J Expo Anal Environ Epidemiol 6:23-34.

Pope CAI, Dockery DW. **2006**. Health effects of fine particulate air pollution: lines that connect. Air Waste 56:709-742.

Preacher KJ, Hayes AF. **2008**. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Meth 40:879-891.

Prentice RL, Sheppard L. **1995**. Aggregate data studies of disease risk factors. Biometrika 82:113-125.

Reis MA, Alves LC, Wolterbeek HTh, Verburg TG, Freitas MC, Gouveia A. **1996**. Main atmospheric heavy metal sources in Portugal by biomonitor analysis. Nucl Instrum Meth B 109-110:493-497.

Reis MA, Alves LC, Freitas MC, van Os B, Wolterbeek HTh. **1999**. Lichens (*Parmelia sulcata*) time response model to environmental elemental availability. Sci Total Environ 232:105-115.

Reis MA. **2001**. *Biomonitoring and assessment of atmospheric trace elements in Portugal - Methods, response modelling and nuclear analytical techniques.* PhD, Delft University of Technology.

Reis MA, Alves LC, Freitas MC, van Os B, de Goeij JJM, Wolterbeek HTh. **2002**. Calibration of lichen transplants considering faint memory effects. Environ Pollut 120:87-95.

Reynolds HD. **1998**. *The modifiable area unit problem: Empirical analysis by statistical simulation.* PhD, University of Toronto.

Roberts S. **2005**. Using moving total mortality counts to obtain improved estimates for the effect of air pollution on mortality. Environ Health Perspect 113:1148-1152.

Robins JM, Morgenstern H. **1987**. The foundations of confounding in epidemiology. Comput Math Appl 14:869-916.

Robinson WS. **2009**. Ecological correlations and the behavior of individuals. Int J Epidemiol 38:337-341.

Roosli M. **2001**. *Spatial variability of air pollutants in the Basel area and carcinogenic and non-carcinogenic health risk.* PhD, University of Basel.

Rose CI, Hawksworth DL. **1981**. Lichen recolonization in London's cleaner air. Nature 289:289-292.

Rose G, Day S. **1990**. The population mean predicts the number of deviant individuals. BMJ 301:1031-1034.

Rosenbaum PR, Rubin DB. **1984**. Difficulties with regression analyses of age-adjusted rates. Biometrics 40:437-443.

Ross H. **1990**. On the use of mosses (*Hylocomium splendens* and *Pleurozium schreberi*) for estimating atmospheric trace metal deposition. Water Air Soil Poll 50:63-76.

Rothman KJ. **2002**. Epidemiology: an introduction. Oxford University Press.

Ruhling A, Tyler G. **1968**. An ecological approach to the lead problem. Botaniska Notiser 121:321-342.

Ruhling A, Tyler G. **1973**. Heavy metal deposition in Scandinavia. Water Air Soil Poll 2:445-455.

Ruhling A. **1994**. *Atmospheric heavy metal deposition in Europe - estimation based on moss analysis*. Copenhagen: Nordic Council of Ministers 9, 1-53.

Rusu AM, Jones GC, Chimonides PDJ, Purvis OW. **2006**. Biomonitoring using the lichen *Hypogymnia physodes* and bark samples near Zlatna, Romania immediately following closure of a copper ore-processing plant. Environ Pollut 143:81-88.

Salway R. **2003**. *Statistical issues in the analysis of ecological studies.* PhD, Imperial College School of Medicine at St. Mary's.

Salway R, Wakefield J. **2004**. A common framework for ecological inference in Epidemiology, Political Science, and Sociology. In: King G, Rosen O, Tanner MA, eds. Ecological inference: New methodological strategies. Cambridge University Press, 303-332.

Samet JM. **2000**. What properties of particulate matter are responsible for health effects? Inhal Toxicol 12:19-21.

Samoli E, Schwartz J, Wojtyniak B, Touloumi G, Spix C, Balducci F, et al. **2001**. Investigating regional differences in short-term effects of air pollution on daily mortality in the APHEA project: a sensitivity analysis for controlling long-term trends and seasonality. Environ Health Perspect 109:349-448.

Sarmento SM, Wolterbeek HTh, Verburg TG, Freitas MC. **2008**. Correlating element atmospheric deposition and cancer mortality in Portugal: Data handling and preliminary results. Environ Pollut 151:341-351.

Schlesinger RB, Kunzli N, Hidy GM, Gotschi T, Jerrett M. **2006**. The health relevance of ambient particulate matter characteristics: Coherence of toxicological and epidemiological inferences. Inhal Toxicol 18:95-125.

Schwartz J. **1994**. Air pollution and daily mortality: a review and meta analysis. Environ Res 64:36-52.

Schwartz J. **2000a**. Assessing confounding, effect modification and thresholds in the association between ambient particles and daily deaths. Environ Health Perspect 108:563-568.

Schwartz J. **2000b**. Harvesting and long-term exposure effects in the relation between air pollution and mortality. Am J Epidem 151:440-448.

Seaward MRD, Letrouit-Galinou MA. **1991**. Lichen recolonization of trees in the Jardin du Luxembourg, Paris. Lichenologist 23:181-186.

Sheppard L. **2003**. Insights on bias and information in group-level studies. Biostatistics 4:265-278.

Sloof JE, Wolterbeek HTh. **1991**. National trace-element air pollution monitoring survey using epiphytic lichens. Lichenologist 23:139-165.

Sloof JE, Wolterbeek HTh. **1992**. Lichens as biomonitors for radiocaesium following the Chernobyl accident. J Environ Radioact 16:229-242.

Sloof JE. **1993**. *Environmental lichenology: Biomonitoring trace element air pollution.* PhD, Delft University of Technology.

Sloof JE, Wolterbeek HTh. **1993a**. Interspecies comparison of lichens as biomonitors of trace-element air pollution. Environ Monit Assess 25:149-157.

Sloof JE, Wolterbeek HTh. **1993b**. Substrate influence on epiphytic lichens. Environ Monit Assess 25:225-234.

Smith RL, Davis JM, Sacks J, Speckman P, Styer P. **2000**. Regression models for air pollution and daily mortality: analysis of data from Birmingham, Alabama. Environmetrics 11:719-743.

Smodis B, Parr RM. **1999**. Biomonitoring of air pollution as exemplified by recent IAEA programs. Biol Trace Elem Res 71-72:257-266.

Smodis B, Bleise A. **2002**. Internationally harmonized approach to biomonitoring trace element atmospheric deposition. Environ Pollut 120:3-10.

Smodis B. **2003**. IAEA approaches to assessment of chemical elements in atmosphere. In: Markert BA, Breure AM, Zechmeister HG, eds. Bioindicators & Biomonitors: Principles, concepts and applications. Elsevier, 875-902.

Smodis B, Bleise A. **2007**. IAEA quality control study on determining trace elements in biological matrices for air pollution research. J Radioanal Nucl Ch 271:269-274.

Steel DG, Beh EJ, Chambers RL. **2004**. The information in aggregate data. Ecological Inference: New methodological strategies. Cambridge University Press, 51-68.

Steinnes E, Hanssen JE, Rambaek JP, Vogt NB. **1994**. Atmospheric deposition of trace elements in Norway: temporal and spatial trends studied by moss analysis. Water Air Soil Poll 74:121-140.

Stine R. **1989**. An introduction to bootstrap methods. Sociol Methods Res 18:243-291.

Subramanian SV, Jones K, Kaddour A, Krieger N. **2009**. Revisiting Robinson: The perils of individualistic and ecologic fallacy. Int J Epidemiol 38:342-360.

Szczepaniak K, Biziuk M. **2003**. Aspects of the biomonitoring studies using mosses and lichens as indicators of metal pollution. Environ Res 93:221-230.

Thurston GD, Ito K, Mar TF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV, Liu H, Neas LM, Pinto J, Stolzel M, Suh H, Hopke PK. **2005**. Workgroup report: workshop

on source apportionment of particulate matter health effects-intercomparison of results and implications. Environ Health Perspect 113:1768-1774.

Tzelgov J, Henik A. **1985**. A definition of suppression situations for the general linear model: A regression weights approach. Educ Psychol Meas 45:281-284.

Tzelgov J, Henik A. **1991**. Suppression situations in Psychological research: definitions, implications, and applications. Psychol Bull 109:524-536.

Vedal S. **1997**. Ambient particles and health: lines that divide. Air Waste 47:551-581.

Verburg TG, Sarmento SM, Wolterbeek HTh. **2010**. Statistical approaches in environmental epidemiology. In: Lahiri S, ed. Advanced Trace Analysis. New Delhi: Narosa Publishing House, 1-69.

Wakefield J, Salway R. **2001**. A statistical framework for ecological and aggregate studies. J R Stat Soc Ser A 164:119-137.

Wakefield J. **2003**. Sensitivity analyses for ecological regression. Biometrics 59:9-17.

Wakefield J. **2004**. A critique of statistical aspects of ecological studies in spatial epidemiology. Environ Ecol Stat 11:31-54.

Wakefield J, Shaddick G. **2006**. Health-exposure modeling and the ecological fallacy. Biostatistics 7:438-455.

Wakefield J, Haneuse SJ-P. **2008**. Overcoming ecologic bias using the two-phase study design. Am J Epidem 167:908-916.

Wakefield J. **2008**. Ecologic studies revisited. Annu Rev Public Health 29:75-90.

Wappelhorst O, Kuhn I, Oehlmann J, Markert BA. **2000**. Deposition and disease: a moss monitoring project as an approach to ascertaining potential connections. Sci Total Environ 249:243-256.

Webster TF. **2007**. Bias magnification in ecologic studies: a methodological investigation. Environ Health 6:1-17.

WHO. **1999**. *Monitoring ambient air quality for health impact assessment.* Copenhagen: World Health Organisation European Series N85,

WHO. **2000**. Air Quality Guidelines for Europe: Second edition. Copenhagen: World Health Organization Regional Office for Europe.

WHO. **2005**. *Air Quality Guidelines. Global Update 2005.* Copenhagen: World Health Organisation

Willis A, Krewski D, Jerrett M, Goldberg MS, Burnett R. **2003**. Selection of ecologic covariates in the American Cancer Society Study. J Toxicol Environ Health A 66:1563-1590.

Wolterbeek HTh, Bode P, Verburg TG. **1996**. Assessing the quality of biomonitoring via signal-to-noise ratio analysis. Sci Total Environ 180:107-116.

Wolterbeek HTh. **2001**. Large-scaled biomonitoring of trace element air pollution: goals and approaches. Radiat Phys Chem 61:323-327.

Wolterbeek HTh. **2002**. Biomonitoring of trace element air pollution: principles, possibilities and perspectives. Environ Pollut 120:11-21.

Wolterbeek HTh, Verburg TG. **2002**. Judging survey quality: Local variances. Environ Monit Assess 73:7-16.

Wolterbeek HTh, Verburg TG. **2004a**. Atmospheric metal deposition in a moss data correlation study with mortality and disease in the Netherlands. Sci Total Environ 319:53-64.

Wolterbeek HTh, Verburg TG. **2004b**. Judging survey quality in biomonitoring. In: Wiersma GB, ed. Environmental Monitoring. CRC Press, 583-603.

Wolterbeek HTh, Sarmento SM, Verburg TG. **2010**. Is there a future for biomonitoring of elemental air pollution? A review focused on a larger-scaled health-related (epidemiological) context. J Radioanal Nucl Ch 286:195-210.

Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery DW, Cohen A. **2000**. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environ Health Perspect 108:419-426.

Zidek JV, Wong H, Le ND, Burnett R. **1996**. Causality, measurement error and multicollinearity in epidemiology. Environmetrics 7:441-451.

# Summary

The context for this thesis is the concern that exposure to environmental air pollution causes adverse health effects in the human population. For the studies presented here data are used on either the Lisbon's population or of about half of the population of continental Portugal. As exposure to air pollution exposure indicators either regulated air pollutants ($PM_{10}$, $SO_2$, etc) are used or a wide variety of chemical elements measured through biomonitoring with lichens. The primary aim of this thesis however, is not to estimate effects from air pollutants but to explore how data and methodological uncertainties can affect results.

Chapter 1 summarises some of the most important and debatable issues in Epidemiology.

Chapter 2 uses time-series data to relate daily air pollutant levels with daily hospital admissions in Lisbon over 5.5 years. Time series studies are very suitable in that they have non-aggregated data (daily over many years). In time-series the unit of analysis is the day, partly because this is the minimum unit for which health data is available and partly because the aim is to investigate acute health effects. It is known that non-aggregated data can be ridden with noise and outliers, which aggregation can help to improve. In many time-series studies robust regression is used in order to deal with outliers in the health data. Chapter 2 shows that using a moving average on both the exposure and health variable, does not affect response estimates but greatly improves their precision and robustness, and to a greater extent than robust regression. This implies that, at least with our data, and for a 7 day moving average information was not lost by aggregation.

Chapter 3 is focused on geographical studies which present the opposite problem of time-series studies, the data is often very aggregated and consequently epidemiological studies may be based on perhaps a small fraction of the total variability in exposures. Chapter 3 addresses how many samples should be taken in order to represent populations with both normal and lognormal distributions at a wide range of exposures. A bootstrap-based method is also presented that enables investigators to simulate the necessary number of samples to represent a population and effect variability. The results show that presently used air pollution exposure data, whether biomonitoring or instrumental monitoring, are insufficient to represent the true variance of the population and especially of the margin of the uncertainty of the sampling survey.

Chapter 4 uses municipality-aggregated data to associate hospital admissions over 11 years and chemical elements measured by biomonitoring. The bootstrap is used to investigate issues such as data uncertainty and model selection uncertainty. Regression errors are shown to be underestimated in about half of the relationships and a-priori model selection is misleading due the inability to account for the uncertainty in model selection. Despite the caveats of this study, including the fact that it is ecological and the likely large within-municipality heterogeneity in exposure to the chemical elements, of these chemical elements can be generally concluded that 1% increase in pollutants may be associated with an average 14% increase in hospital admissions.

Chapter 5 uses the same data as Chapter 4 to investigate the prevalence of suppression, also known as negative confounding, in a rather general dataset. Identification of the specific confounding situation, that affects each relationship is important because it can help in the eventual interpretation of outcomes and because suppression can lead to too optimistic end results in terms of errors and model fit, in the same way that the more known positive confounding leads to too pessimistic errors and model fit. Chapter 5 shows that suppression is quite common, and affects about 35% of the studied relationships. Chapter 5 also presents a more detailed and clear description of possibilities to distinguish suppression.

Chapter 6 provides an overall summary and conclusions for the thesis.

# Samenvatting

De context van dit proefschrift is de bezorgdheid dat blootstelling aan omgevingsluchtverontreiniging negatieve gezondheidseffecten kan veroorzaken in de humane populatie. Van de in dit proefschrift aangehaalde studies zijn data gebruikt van zowel de populatie van Lissabon als van ongeveer de helft van de populatie van continentaal Portugal. Voor de blootstelling aan luchtverontreiniging zijn ofwel gereguleerde luchtverontreigingen gebruikt (PM10, SO2 enz), ofwel een grote variatie aan chemische elementen, gemeten via biomonitoring met behulp van korstmossen. Het primaire doel van dit proefschrift is echter niet de bepaling van de effecten van luchtverontreinigende stoffen maar het nagaan van hoe gegevens en methodologische onzekerheden van invloed kunnen zijn op resultaten.

Hoofdstuk 1 vat enkele items samen die als het meest belangrijk zijn opgevat en beschouwd als discutabel binnen de epidemiologie.

Hoofdstuk 2 gebruikt gevens in tijd-series waarin luchtverontreinigingen worden gerelateerd aan dagelijkse ziekenhuisopnames in Lissabon, over een periode van 5,5 jaar. Tijd-serie studies zijn zeer geschikt omdat zij bestaan uit niet-geaggregeerde gegevens (dagelijks, over vele jaren). In tijd-series, de dag is de analytische unit, gedeeltelijk omdat dat de minimale unit is waarvoor gezondheidsdata beschikbaar zijn, en gedeeltelijk omdat het doel is om acute gezondheidseffecten te onderzoeken. Het is bekend dat niet-geagregeerde data behept kunnen zijn met ruis en uitschieters, wat via aggregatie verbeterd kan worden. In veel tijd-serie studies wordt robuuste regressie gebruikt om om te gaan met uitschieters in de gezondheidsdata. Hoofdstuk 2 laat zien dat gebruikmaking van "bewegende gemiddelden" voor zowel de blootstelling als de gezondheidsvariabele geen effect heeft op de responseschattingen, maar dat hun precisie en robuustheid sterk verbeterd wordt, dit in een grotere mate dan via robuuste regressiemethoden. Dit houdt in dat, in ieder geval met de gebruikte data, en voor een 7-daags "bewegend gemiddelde", geen informatie verloren raakt als gevolg van aggregatie.

Hoofdstuk 3 is voornamelijk gericht op geografische studies die een tegenovergeteld probleem inhouden van de tijd-series studies: de gegevens zijn veelal in verregaande mate geaggregeerd en als gevolg daarvan kunnen epidemiologische studies gebaseerd zijn op slechts een kleine fraktie van de totale variabiliteit in blootstelling. In hoofstuk 3 wordt ingegaan op de hoeveelheid samples die nodig zijn om populaties te representeren met normale- en lognormale distributies ten aanzien van een grote verscheidenheid aan blootstellingen. Een methode gebaseerd op boot-strapping is gepresenteerd die het onderzoekers mogelijk maakt om de noodzakelijke hoeveelheid samples te simuleren om een populatie (en effect-variatie) adekwaat te representeren. De resultaten laten zien dat de huidig gebruikte gegevens ten aanzien de blootstelling aan luchtverontreiniging, of het nu gaat om biomonitoring of instrumentele monitoring, ontoereikend zijn om de werkelijke variatie in de effecten te weerspiegelen, en hierbij met name de omvang van de fout in de sampling survey.

Hoofdstuk 4 gebruikt gemeente-geagregeerde gegevens om ziekenhuisopnames over 11 jaar te associeren met data ten aanzien van chemische elements verkregen uit biomonitoring. Bootstrapping is gebruikt om gegevens te verkrijgen ten aanzien van onzekerheden in gegevens en ten aanzien van de te selecteren modellen. Onzekerheden in regressies blijken te worden onderschat in ongeveer de helft van alle gehanteerde relaties, en a-priori selectie in te hanteren modellen is misleidend in die zin dat het onmogelijk lijkt om daarvan de resulterende onzekerheden in te schatten. Ondanks het voorbehoud bij deze studie, waaronder het feit dat het gaat om een ecologische studie en de waarschijnlijk grote gemeentelijke heterogeniteiten in blootstelling aan de chemische stoffen, kan voor deze chemische stoffen in algemene zin worden gezegd dat 1 % toename in verontreiniging kan worden geassocieerd met een gemiddeld 14 % toename in ziekenhuisopname.

Hoofdstuk 5 gebruikt dezelfde gegevens als hoofdstuk 4 om de prevalentie van suppressie te onderzoeken, ook bekend als negatieve confounding (verstoring), dit in een doorsnee dataset. Identificatie van de specifieke confounding situatie die van invloed is op elke relatie is belangrijk omdat dit kan helpen in de uiteindelijke interpretatie van uitkomsten, en omdat suppressie kan leiden tot te positieve eindresultaten voor wat betreft zowel fouten als model-fits op dezelfde wijze als dat de meer bekendere positieve confounding kan leiden tot te negatieve interpretatie van fouten en model-fit. Hoofdstuk 5 geeft aan dat suppressie een algemeen verschijnsel is, en van invloed op ongeveer 35 % van alle bekeken relaties. Hoofdstuk 5 geeft ook een meer gedetailleerde en duidelijke omschrijving van mogelijkheden om suppressie te onderscheiden.

Hoofdstuk 6 geeft een algemene samenvatting en conclusies van het proefschrift.

# Acknowledgments

I was extremely fortunate to have two amazing supervisors. Prof. Bert Wolterbeek whose fabulous memory, drawings worth of Pratt, and a kind of sophisticated naivety take on science was an unflickering guiding light. Prof. Maria do Carmo Freitas the tireless multi-tasker whose persistence and pragmatism were as indispensable as her friendship. Thank you!

Quite literally, I was standing on the shoulders of giants to be able to accomplish this work. Tona Verburg my little big (and vice-versa) girl, from MatLab scribbles to travelling to the last row x column of excel, from sulphur enriched dressings of semi-carbonised protein to ultrasounds, and that problems do not exist, they are made… Thank you so much for your friendship, come back! You know you have a PhD honoris causa. The Future is thus closed.

I wish to show my appreciation to researchers whose work was systematically inspiring and critical for this thesis: Dr. Frederick W. Lipfert, Prof. Jon Wakefield, Dr. Ruth Salway and Prof. Lianne Sheppard. I also wish to thank the amazing altruistic project that is the evolving online textbook of Epidemiology by Dr. Victor Schoenback.

This thesis would not have been at all possible without the colossal effort invested on the part of the data providers in Portugal. The Instituto Nacional de Estatistica (INE) and the Administração Central do Sistema de Saúde (ACSS) provided the health databases for which I am incredibly thankful. I particularly wish to thank the advice and assistance of Dr. Luzia Estevens (INE) and of Dr. Teresa Boto (ACSS). In addition, the Portuguese Environmental Agency, National Institute of Meteorology and the Military Geographical Institute were instrumental in providing additional environmental data and introducing me to the distorted world of projections and datums. I would also like to thank Dr. Miguel Reis for helping revising the original biomonitoring database. Ramon thank you for those precious ArcGIS shape files that were impossible to get and for introducing me to Utrecht and the amazing world of miniatures.

The picture on the cover of this book is called "Sea of Lines" and was created by Jean-Nöel Lafargue. Thank you for sharing so trustingly.

Andreas you helped me when I needed the most. You, Albina, Alexander and Lena were my family environment in Delft. Our talks were always like the first spring day after a long winter: enlightened blossoming sparkling adventurous revolutionary. You have no idea how I miss them. From the cocoon trees of Rijswijk to the story of the man with the white paper, thank you.

Ivo from chemical politics to political chemistry, from orange antidotes to francesinhas crude oil, your wide knowledge and gentle intelligence are a rare combination. Always park overground and never fly downstairs. See you in the next industrial plant.

Marnix. Confucian metallican salsacian. Could hardly see you with your busy agenda, but when I did I was treated to risottos, PET talk and private live music upon request. You lent

me Zen and the Art of Volkswagen Maintenance and the Hitchhiker's Guide to German Highways. It was hilarious!

Auréle it was remarkable to find out that one needs friction to get one's moving, that terror can be tender, and that decadence can be sweet, that inexistent words can be meaningful.

Karoly living upstairs from a construction site we spent that Carnival listening to Dutch pop music while locals prepared for war. Thank you for the wonderful Hungarian hospitality and bohemia.

Baukje, the Molybdenum tom-tom. I will never forget when we realised we had missed the last boat and how the extra 3hours (was it?) of cycling did not prevent us from delighting on those wonderful Belgian fries. How come all days I remember with you were sunny? Was it that yellow hair?

Candice, the moving rainbow target. On that hellish spiral that is to get to/from Utrecht at night, on a vulcanised bike, a porcelain violent femme, you lived in a permanent construction site and worked in a mud spattered laboratory. You are THE Chicago's manual of style. And your husband is not too bad either.

Jaap Jan, la resistance of the ice. There is a lot of latent heat in ice. Your light seriousness and unbearable lightness were spiritual tonics. Hope that works on the dykes. Never read your books; I am saving them for a special occasion. Thank you for all the crying laughs.

Antonia, the driver from the Black Sea. Your gentle kindness and sort of Bulgakovian sense of the absurd and of the sensible were one of the most immediate things I noticed in you. I will take a lift from you anytime, especially to tango.

Jeroen, the neutron volley. Sabotaging poker games, conspiring jam sessions, whirling page numbers, the house in New Orleans, what is it with you? Just a great friend I guess, unconditional. That book you gave me, it was the first and last time I understood Dutch from start to finish.

Sander, those days of transmuting metal and rock was magic. Your reciting of that Jane Austen's passage was poetry. And that mount the hard-drive software was a life saver.

Niels your tolerance and free spirit are inspiring, it's that simple. Heidi thanks for reminding me that personalities can be so much more colourful so much more passionate, and it's so so so nice!

Dennis, it is just unusual and awesome to share the liking of both electronic and jazz music. That undecipherable handwriting and miniature computer are unlike the sound of your drums.

Diane, allergic to her chemical reaction, thank you for that crazy rugby game and all the southern rendez-vous. Enrica and David thank you so much for the good times together. Dan, Nicole and Matt, coming from that twilight place that is the first American ruins (you know I mean it well), you landed Lisboa which is not in a good state either. Hope to hear good country music and play that card game again with you guys.

John you were one of the few native English speakers but the only one person to always address me in Dutch. Thanks for the reminiscences of England and the sodium polyacrylate stories from the basement. Gustavo thank you for that non-stop tango and Jose thank you for that tortilla, still would like the recipe. A-ha-lastair, I remember you trying to boil a shell-less egg in the microwave, how did that go? Romee, the long-haul flamenco cyclist, I hope you are now healthily less fit and not facing the narrow abyss of de Hague's tram trails. Marc still thinking of listening to that saxophone. Leon, Lucas, Wouter, the famed gamblers, thank you for spoiling the evenings.

Around a flying carpet parked on top of a table, which rests on top of a sedentary carpet, meet some very special people. Yvon an uncooperative life rescuer who never gets peace and quiet, thank you so much for every single one of those little and big things you handled for me. Folkert the surgeon of my computer, you treated fire, viruses, trojans and frogs, you did transplantations, prosthetics, kryogenics, and scrambled eggs; you always fixed it. Delia the toxic woman whose feet don't reach the ground, thank you for the laughs. Astrid thank you for the nice stories about Bert (one involving a bike and a ditch) the laughter from beyond and for setting those mice free. Anneke thank you for the fast pneumatic system competition and all the work we did together on solving that salty diapers problem. Thea thank you for the advice on the geology of soils and that lovely dinner in Delft. Mehmet thank you for just being really cool and easy-going and for having worked on some of the most interesting applications of INAA on commercial products. Jan Willem, the molybdenum trainspotter, thanks for your wise company. Helleen, Zvonko, Olav, Peter and Ulla thank you for all those coffee breaks. Klazien, Mark, Tineke and Marcel thank you for the twilight insights into nuclear physics/chemistry.

Henk and Koos may I have a refill, in small doses? Thank you so much for your support in those long experiments.

Joana minha querida obrigado por me receberes tão bem, pela comida, pela dança, pelos desabafos. Lenie thank you so much for spontaneously getting me that box and for the attempts at making me understand and speak Dutch, it was a hopeless endeavour but it was a funny one.

I wish to acknowledge the company and support for some very early PhD students. David thank your for the lessons in Dostoyevsky and speed printing. Jessica thank you for the LC concert. Alexander and Lambert thanks for showing me around Delft. Erica thank your for the talks. I also wish to thank everybody in de Clok, including Roeland, Gabi and Jet and…I don't remember your names, but I am sure you understand.

To my colleagues at ITN: Rita thanks for being a good-sports when it comes to sharing homes with students and for all the support from your thesis which was very important for mine. Mané you are incredibly brave when it comes to cockroaches maybe it's because you have cat eyes! I am waiting for a line dedicated to me on your book. Ana Paula, I guess you will never forget to change the time again (nor Tona), thanks for the cool and your wonderful singing. Bruno all the best to you out there in the sea. I also wish to thank others at ITN. Marta the

choreographer charged with somebody else's fines; Nuno the actor/writer in need of horns or a helmet. Alexandra the homeopathic chocolate provider. Ana the space-shuttle that makes me laugh. Marina and Catarina thanks for your kindness. Ho Manh Dung from whom I learnt INAA and detector calibration, you are a true and dedicated scientist. Also at ITN, Joana you helped me like Noah's arc, and Teresa thanks for your permanent impossible high-spirits and diligence.

I was very fortunate to have found an exquisite group of people practicing my favourite sports. Thank you to our formidable teacher Xialin and to my dear colleagues: Sarah, Ping, Aura and Franciska.

I am incredibly indebted to Niels, for taking me in, despite my manic need for a pristine kitchen. There is nothing that you cannot do from operating farming machines to dismantling cars and walls, you are a great guy. Tina you will be a great engineer.

Aos meus amigos em Portugal. Sandra pela tua companhia ao longo destes anos, apesar dos altos e baixos. Bruno, de pedaços de chocolates ao mi menor, de Yes aos sistemas dinâmicos, teoremas complexos, cabelos caóticos, encontramo-nos no próximo strange attractor, nem que seja uma passadeira. Rita, Catarina, Carlos, Daniela, Ana, João, Paulo obrigado pela amizade ao longo destes muitos anos e dos que mais virão.

As palavras sempre falham quando são mais necessárias, mais merecedoras. Para a minha irmã Elsa, a minha mãe e o meu pai, que foram tão mais do que o imaginável. Um grande OBRIGADO a todos por me aturarem, e para o António também. Inês e Francisco que possam lembrar tudo o que sonham agora em crianças, e alcançá-lo. Uma chuva de beijinhos resmungões.

# Curriculum Vitae

Susana F. M. Sarmento was born on the 25th March 1978 in Porto, Portugal. She obtained her high school degree at Escola Secundária Lumiar 1 in Lisbon. Subsequently she studied Biology at the University of York in England with a dissertation entitled "Pre-conditioned feeding behaviour in slugs". Later she obtained a Masters in Neuroscience at the University of Edinburgh in Scotland with a thesis on the "Sterol and steroid metabolism in human brain tumours". She started a PhD at the RID of the Delft University of Technology the results of which are presented in this book.

# List of Publications

Sarmento SM, Verburg TG, Freitas MC, Wolterbeek HTh. **Submitted January 2012** *to Inhalation Toxicology*. Geographical association of airborne metals, measured with biomonitoring, with cardiovascular disease in the Portuguese population.

Sarmento SM, Verburg TG, Freitas MC, Wolterbeek HTh. **Submitted January 2012** *to Inhalation Toxicology*. Suppression situations in the geographical association between airborne metals and cardiovascular disease – application and implications for environmental epidemiology.

Sarmento SM, Verburg TG, Almeida SM, Freitas MC, Wolterbeek HTh. **2011**. Robustness of different regression modelling strategies in epidemiology: a time-series analysis of hospital admissions and air pollutants in Lisbon (1999-2004). Environmetrics 22:86-97.

Wolterbeek HTh, Sarmento SM, Verburg TG. **2010**. Is there a future for biomonitoring of elemental air pollution? A review focused on a larger-scaled health-related (epidemiological) context. J Radioanal Nucl Ch 286:195-210.

Verburg TG, Sarmento SM, Wolterbeek HTh. **2010**. Statistical approaches in environmental epidemiology. In: Lahiri S, ed. Advanced Trace Analysis. New Delhi: Narosa Publishing House, 1-69.

Sarmento SM, Wolterbeek HTh, Verburg TG, Freitas MC. **2008**. Correlating element atmospheric deposition and cancer mortality in Portugal: Data handling and preliminary results. Environ Pollut 151:341-351.

Dung H, Freitas MC, Sarmento SM, Blaauw M, Beasley D. **2008**. Calibration of gamma-ray spectrometers coupled to Compton suppression and fast pneumatic systems for the $k_0$-standardized NAA method. J Radioanal Nucl Ch 278:621-625.

Freitas MC, Pacheco AMG, Dionisio I, Sarmento SM, Baptista MS, Vasconcelos MTSD, Cabral JP. **2007**. Instrumental neutron activation analysis and inductively coupled plasma mass spectrometry on atmospheric biomonitors. J Radioanal Nucl Ch 273:705-711.

Pacheco AMG, Freitas MC, Sarmento SM. **2007**. Nuclear and non-nuclear techniques for assessing the differential uptake of anthropogenic elements by atmospheric biomonitors. Nucl Instrum Meth A 579:499-502.

# Appendix

Sample sizes needed for representing the population average, determined by simulation

| | | Normal distribution | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. |
| | | 9.96 | 25.14 | -0.04 | 0.06 | 10.03 | 49.37 | 0.01 | 0.06 | 10.01 | 98.58 | -0.02 | 0.09 | 9.82 | 150.47 | 0.01 | -0.05 |
| | | margin of error | | | | margin of error | | | | margin of error | | | | margin of error | | | |
| | | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 |
| Significance | 85 | 6 | 9 | 19 | 72 | 19 | 32 | 73 | 259 | 73 | 131 | 256 | 744 | 165 | 283 | 528 | 1175 |
| | 90 | 8 | 11 | 23 | 89 | 23 | 40 | 89 | 308 | 89 | 153 | 301 | 844 | 194 | 340 | 616 | 1268 |
| | 95 | 9 | 14 | 30 | 117 | 30 | 52 | 111 | 376 | 114 | 194 | 405 | 964 | 261 | 432 | 721 | 1398 |
| | 99 | 13 | 22 | 43 | 182 | 47 | 82 | 178 | 514 | 168 | 280 | 549 | 1181 | 375 | 566 | 935 | 1556 |
| | | Log-normal distribution | | | | | | | | | | | | | | | |
| | | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. |
| | | 10.03 | 25.24 | 0.69 | 0.51 | 9.96 | 49.91 | 1.92 | 8.19 | 10.14 | 112.61 | 6.60 | 93.22 | 10.22 | 153.82 | 7.20 | 98.68 |
| | | margin of error | | | | margin of error | | | | margin of error | | | | margin of error | | | |
| | | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 |
| Significance | 85 | 6 | 10 | 20 | 72 | 18 | 31 | 68 | 259 | 88 | 150 | 321 | 876 | 150 | 254 | 512 | 1143 |
| | 90 | 7 | 11 | 24 | 86 | 23 | 38 | 83 | 307 | 109 | 181 | 367 | 976 | 176 | 300 | 594 | 1234 |
| | 95 | 9 | 15 | 31 | 116 | 29 | 51 | 111 | 403 | 137 | 230 | 474 | 1089 | 237 | 370 | 701 | 1349 |
| | 99 | 12 | 22 | 49 | 178 | 51 | 73 | 162 | 549 | 220 | 354 | 663 | 1275 | 326 | 513 | 896 | 1574 |

Sample sizes needed for representing the population variance, determined by simulation

| | | Normal distribution | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. |
| | | 9.96 | 25.14 | -0.04 | 0.06 | 10.03 | 49.37 | 0.01 | 0.06 | 10.01 | 98.58 | -0.02 | 0.09 | 9.82 | 150.47 | 0.01 | -0.05 |
| | | margin of error | | | | margin of error | | | | margin of error | | | | margin of error | | | |
| | | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 |
| Significance | 85 | 39 | 68 | 145 | 474 | 38 | 65 | 146 | 494 | 39 | 67 | 148 | 494 | 35 | 60 | 136 | 455 |
| | 90 | 48 | 82 | 174 | 536 | 47 | 78 | 175 | 558 | 47 | 85 | 181 | 551 | 42 | 72 | 161 | 510 |
| | 95 | 62 | 106 | 224 | 705 | 59 | 100 | 220 | 677 | 60 | 117 | 233 | 674 | 54 | 98 | 203 | 627 |
| | 99 | 94 | 153 | 348 | 935 | 90 | 159 | 325 | 852 | 104 | 176 | 313 | 890 | 79 | 138 | 285 | 860 |
| | | Log-normal distribution | | | | | | | | | | | | | | | |
| | | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. |
| | | 10.03 | 25.24 | 0.69 | 0.51 | 9.96 | 49.91 | 1.92 | 8.19 | 10.14 | 112.61 | 6.60 | 93.22 | 10.22 | 153.82 | 7.20 | 98.68 |
| | | margin of error | | | | margin of error | | | | margin of error | | | | margin of error | | | |
| | | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 |
| Significance | 85 | 44 | 80 | 173 | 554 | 143 | 251 | 510 | 1188 | 772 | 1080 | 1596 | 1710 | 823 | 1148 | 1601 | 1735 |
| | 90 | 54 | 95 | 201 | 639 | 175 | 298 | 604 | 1264 | 826 | 1183 | 1720 | 1785 | 887 | 1284 | 1736 | 1822 |
| | 95 | 68 | 119 | 267 | 753 | 240 | 383 | 709 | 1432 | 921 | 1379 | 1872 | 1902 | 1004 | 1472 | 1876 | 1914 |
| | 99 | 106 | 186 | 389 | 1003 | 348 | 528 | 922 | 1667 | 1182 | 1669 | 1986 | 1987 | 1285 | 1710 | 1988 | 1988 |

Sample sizes needed for representing the average of an infinite population, determined by StatGraphics

| | | Normal distribution | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. |
| | | 9.96 | 25.14 | -0.04 | 0.06 | 10.03 | 49.37 | 0.01 | 0.06 | 10.01 | 98.58 | -0.02 | 0.09 | 9.82 | 150.47 | 0.01 | -0.05 |
| | | margin of error | | | | margin of error | | | | margin of error | | | | margin of error | | | |
| | | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 |
| Significance | 85 | 5 | 8 | 15 | 54 | 15 | 25 | 53 | 204 | 52 | 91 | 203 | 803 | 114 | 202 | 452 | 1799 |
| | 90 | 7 | 10 | 20 | 71 | 19 | 32 | 68 | 266 | 68 | 119 | 264 | 1048 | 149 | 263 | 590 | 2349 |
| | 95 | 9 | 14 | 27 | 100 | 26 | 45 | 97 | 377 | 96 | 168 | 375 | 1487 | 211 | 373 | 834 | 3335 |
| | 99 | 15 | 23 | 46 | 172 | 45 | 76 | 166 | 647 | 165 | 290 | 643 | 2569 | 364 | 640 | 1440 | 5760 |

Sample sizes needed for representing the variance of an infinite population, determined by StatGraphics

| | | Normal distribution | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. | Aver. | CV | Skewn. | Kurt. |
| | | 9.96 | 25.14 | -0.04 | 0.06 | 10.03 | 49.37 | 0.01 | 0.06 | 10.01 | 98.58 | -0.02 | 0.09 | 9.82 | 150.47 | 0.01 | -0.05 |
| | | margin of error | | | | margin of error | | | | margin of error | | | | margin of error | | | |
| | | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 | 0.8-1.2 | 0.85-1.15 | 0.9-1.1 | 0.95-1.05 |
| Significance | 85 | 40 | 65 | 130 | 465 | 40 | 65 | 130 | 465 | 40 | 65 | 130 | 465 | 40 | 65 | 130 | 465 |
| | 90 | 51 | 82 | 167 | 602 | 51 | 82 | 167 | 602 | 51 | 82 | 167 | 602 | 51 | 82 | 167 | 602 |
| | 95 | 70 | 114 | 234 | 849 | 70 | 114 | 234 | 849 | 70 | 114 | 234 | 849 | 70 | 114 | 234 | 849 |
| | 99 | 118 | 193 | 398 | 1455 | 118 | 193 | 398 | 1455 | 118 | 193 | 398 | 1455 | 118 | 193 | 398 | 1455 |

**Stellingen** behorende bij het proefschrift

"Toepassing van atmosferische biomonitoring in de epidemiologie: aandachtspunten ten aanzien van datakwaliteit, monstername, aggregatie en confounding"

Susana F.M. Sarmento

1. Er is geen statistische techniek die goede data kan vervangen. Bijvoorbeeld, geaggregeerde data leveren betrouwbaarder schattingen op dan robuuste regressie. *Dit proefschrift (Hoofdstuk 2)*

2. Veel significante associaties in luchtverontreinigings-epidemiologisch onderzoek betreffen waarschijnlijk suppressieve situaties: de luchtverontreinigingen hebben geen directe relatie met ziekte; zij dragen slechts bij aan de verklaring van de foutvariantie in de confounders. *Dit proefschrift (Hoofdstuk 4)*

3. Van alle wetenschapsgebieden die betrekking hebben op humane populaties (epidemiologie, sociologie, economie, politiek and rechten), zijn rechten, economie and politiek de enige aan wie het toegestaan is te experimenteren met humane populaties, en is epidemiologie de enige vanuit het principe van voorzorg.

4. Het effect van vele tegenwoordig geaccepteerde risicofactoren zoals luchtverontreiniging is waarschijnlijk niet veel groter dan dat van hun nocebo-effect. Zoals in de farmaceutische industrie het effect van een nieuw medicijn groter moet zijn dan dat van de placebo, zou het effect van een negatieve risicofactor groter moeten zijn dan dat van de nocebo.

5. Hoe meer men weet over statistiek, hoe meer men dit weet te manipuleren. Bijvoorbeeld, het recept om je p-waarden te verlagen: vergroot $N$[1,2], vergroot de variantie in $X$[2], voeg variabelen toe of verwijder deze[3], selecteer datapunten[4] en probeer precisieverhogende procedures zoals ridge regressie.*1 Lehrer J. The truth wears off. 13th December 2010. New Yorker, 2 Lipfert F.W. (1999) Journal of Environmental Medicine, 1: 267-278, 3 Breiman L. (1992) Journal of the American Statistical Association, 87 (419): 738-754, 4 Vul E., Harris C., Winkielman P. & Pashler H. (2009) Perspectives on Psychological Science, 4(3): 274-290.*

6. De enige manier om iets echt te begrijpen is verstoring en observatie. *http://www.wired.com/magazine/2011/12ff_causation/, Kacew, S., Toxicology 160, 87-96 (2001)*

7. Als je iets niet uit de hand kunt berekenen (in een spreadsheet om het minder moeilijk te volgen te maken), reken het dan maar helemaal niet uit.

8. Iedere noodzakelijkheid die beantwoord wordt levert tenminste twee nieuwe op die beantwoord moeten worden

9. Een burka is net zo shockerend en onaangepast als een bikini, we zijn alleen gewend geraakt aan de een en niet aan de ander

10. "Informatie is geen kennis, Kennis is geen wijsheid, Wijsheid is geen waarheid,
Waarheid is geen schoonheid, Schoonheid is geen liefde, Liefde is geen muziek,
en Muziek is HET BEST."
Frank Zappa

*Deze stellingen worden opponeerbaar en verdedigbaar geacht en zijn als zodanig goedgekeurd door de promotor(en) Prof. dr. H. Th. Wolterbeek and Prof. dr. M. C. Freitas.*

**Propositions** accompanying the thesis

"Application of Atmospheric Biomonitoring to Epidemiology:

Issues in Data Quality, Sampling, Aggregation and Confounding"

Susana F. M. Sarmento

1. No statistical technique can replace good data. For example, aggregate data provides more reliable estimates than robust regression. *This thesis (Chapter 2)*

2. Many significant associations in current air pollution epidemiological studies are probably suppression situations: the air pollutants have no direct association with disease; they only contribute to explaining the error variance in the confounders. *This thesis (Chapter 4)*

3. Of all sciences dealing with human populations (epidemiology, sociology, economics, politics and law), law, economics and politics are the only ones entitled to perform experiments on human populations, and epidemiology is the only one with a precautionary principle.

4. The effect of many currently accepted risk factors, such as air pollution, is probably not much greater than their nocebo effect. As in the pharmaceutical industry, where the effect of a new medicine must surpass that of the placebo, adverse risk factors should surpass the nocebo effect.

5. The more one knows statistics, the more one knows how to manipulate it. For instance, recipe to decrease your p-values: increase $N^{1,2}$, increase X variance[2], add/remove variables[3], select data points[4] and try precision enhancing procedures such as ridge regression.
*1 Lehrer J. The truth wears off. 13th December 2010. New Yorker, 2 Lipfert F.W. (1999) Journal of Environmental Medicine, 1: 267-278, 3 Breiman L. (1992) Journal of the American Statistical Association, 87 (419): 738-754, 4 Vul E., Harris C., Winkielman P. & Pashler H. (2009) Perspectives on Psychological Science, 4(3): 274-290.*

6. The only way to truly understand something is to interfere and observe.
*http://www.wired.com/magazine/2011/12ff_causation/, Kacew, S., Toxicology 160, 87-96 (2001)*

7. If you can't calculate it by hand (in a spreadsheet to make it less fastidious), don't calculate it.

8. Living up to every necessity generates at least two new ones to live up to.

9. A burka is as shocking and awkward as a bikini, we just got used to one but not the other.

10. "Information is not knowledge, Knowledge is not wisdom,
Wisdom is not truth, Truth is not beauty, Beauty is not love, Love is not music,
and Music is THE BEST."
Frank Zappa

*These propositions are regarded as opposable and defendable, and have been approved as such by the supervisors Prof. Dr. H. Th. Wolterbeek and Prof. Dr. M. C. Freitas.*

ISBN 978-1-61499-048-2

TU Delft

Delft University of Technology

Delft University Press is an imprint of IOS Press

IOS Press