

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

A survey on ecological regression for health hazard associated with air pollution

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Bruno, F., Cameletti, M., Franco Villoria, M., Greco, F.P., Ignaccolo, R., Ippoliti, L., et al. (2016). A survey on ecological regression for health hazard associated with air pollution. SPATIAL STATISTICS, 18(Part A), 276-299 [10.1016/j.spasta.2016.05.003].

Availability:

This version is available at: <https://hdl.handle.net/11585/570533> since: 2016-11-25

Published:

DOI: <http://doi.org/10.1016/j.spasta.2016.05.003>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Bruno, F., Cameletti, M., Franco-Villoria, M., Greco, F., Ignaccolo, R., Ippoliti, L., Valentini, P., Ventrucci, M., 2016. A survey on ecological regression for health hazard associated with air pollution. *Spat. Stat.*, 18, 276–299. <https://doi.org/10.1016/j.spasta.2016.05.003>

The final published version is available online at:

<https://doi.org/10.1016/j.spasta.2016.05.003>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

A Survey on Ecological Regression for Health Hazard Associated with Air Pollution

Francesca Bruno^a, Michela Cameletti^b, Maria Franco-Villoria^c, Fedele Greco^a, Rosaria Ignaccolo^{c,*}, Luigi Ippoliti^d, Pasquale Valentini^d, Massimo Ventrucci^a

^aDept. of Statistical Sciences "Paolo Fortunati", University of Bologna

^bDept. of Management, Economics and Quantitative Methods, University of Bergamo

^cDept. of Economics and Statistics "Cognetti de Martiis", University of Torino

^dDept. of Economics, University of Chieti-Pescara

Abstract

In the last 30 years, a large number of studies have provided substantial statistical evidence of the adverse health effects associated with air pollution. Statistical literature is very rich and includes a plethora of models to manage different types of spatial data. This paper starts with a thorough discussion on the spatial nature of the available data on health and air pollution. Health data are usually provided by Health Authorities as mortality and morbidity counts at a small area level. Thus we mainly focus on reviewing and discussing the spatial and spatio-temporal regression models proposed for disease count data on irregular lattices. In general, measuring the effect of exposure on health outcomes is an extremely hard task, and to obtain reliable estimates of the exposure effect and associated uncertainty one needs to build models that account for the residual variability not captured by the exposure-response relationship. In this context, Bayesian hierarchical models including spatial random effects play a prominent role: we consider both univariate and multivariate models and discuss some extensions to the spatio-temporal setting. Since model estimation can be prohibitive, practitioners are provided with a list of available software for Bayesian inference that avoids the need for complicated coding.

Keywords: Spatial and Spatio-Temporal Regression, Quantitative Health Risk Assessment, Change of Support Problem, Gaussian Markov Random Field, Hierarchical Bayesian Models, Factorial Models

1. Introduction

A substantial literature exists, starting from the late 1980s, on the adverse health effects associated with exposure to high levels of air pollution (see *e.g.* Schwartz and Marcus 1990; Dockery et al. 1993; Brunekreef and Holgate 2002; Schikowski et al. 2005; Lanki et al. 2006; Brook et al. 2010; Künzli 2012; Atkinson et al. 2014; Rushworth et al. 2014). This scientific evidence escalated at the end of 2013 with a study by the International Agency for Research on Cancer¹ which found sufficient evidence to classify air pollution as a leading environmental cause of cancer deaths. The deleterious impact of air pollution on human health is also confirmed by the worldwide 3.7 million premature

*Corresponding author, email: rosaria.ignaccolo@unito.it

¹https://www.iarc.fr/en/media-centre/iarcnews/pdf/pr221_E.pdf

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

deaths attributable to ambient air pollution in 2012, an estimate provided by the World Health Organization (WHO) in 2014². Air pollution is not only a major environmental risk to health but it represents also a societal cost: WHO reports that in 2010 the overall annual economic cost of health impacts and mortality from air pollution, including morbidity costs, amounted to 1.575 trillion US dollars for the countries of the WHO European region (WHO, 2015).

The range of adverse effects of ambient air pollution on health is wide, going from respiratory and cardiovascular diseases to hypertensive disorders and neurodegeneration, as documented by many studies. The impact of air pollution is known to be different depending on the length of the exposure periods, and a significant number of epidemiological studies have been conducted for investigating both the short and long-term effects of exposure. Starting from the 1990s, multicity time series studies have provided evidence of a strong association between short-term exposure to air pollution and adverse health events (in terms of mortality and morbidity). For example, the *National Morbidity, Mortality, and Air Pollution Study* (NMMAPS) in the US and the *Air Pollution and Health: A European Approach* (APHEA) study in Europe have found a 0.21% and a 0.6% increase in total mortality per 10 units elevation in particulate concentrations, respectively (see *e.g.* Samoli et al., 2003 and Peng et al., 2005). From the statistical point of view, this kind of associations are found by using Poisson time series regression models with the daily number of deaths/hospitalizations as outcome, the (possibly lagged) daily level of pollution as a linear predictor and smooth functions of weather variables and calendar time used to adjust for time-varying confounders (see *e.g.* Peng et al., 2006, 2009). There is also a vast literature that considers long-term air pollution exposure (see *e.g.* Carey et al., 2013; Cesaroni et al., 2013; Hoek et al., 2013; Lepeule et al., 2012): in this case data are collected through cohort studies and usually Cox proportional hazards models are used to investigate the associations between pollution concentrations and subsequent cause-specific mortality.

Recent advances in Geographical Information Systems (GIS) and global positioning systems enable accurate geocoding of locations where scientific data are collected. This has encouraged the formation of large monitoring networks to collect exposure measurements. Additionally, pollution concentrations estimated by computer dispersion models on a regular grid have also become widely available in recent times together with satellite measurements of remote sensing. These data sources, featured by a fine spatial resolution, have enabled researchers to examine relationships between disease rates for geographical areas and exposure to environmental risk factors using tools from spatial epidemiology (Elliott and Wartenberg, 2004). In general, health data are continuously collected by Health Authorities and usually consist of mortality and morbidity counts at the small area level, such as irregularly shaped administrative units. On the other hand, measurements of pollutants are associated with a set of monitoring stations and come in the form of geostatistical (or point-referenced) data.

In this review paper, we focus on these types of spatial data and provide a critical review of the statistical methods suggested for the analysis of aggregated count data in the context of ecological spatial and spatio-temporal regression. In particular, we discuss some important critical issues occurring in spatial epidemiology analysis and which typically are related to: *i*) the intrinsic features of the data, which are available from different sources and can be measured with error and with different spatial resolutions; *ii*) the specification of a regression function that typically relates a change in air pollution to health disease risk; *iii*) the acknowledgement of spatial and temporal dependence in regression models, together with potential biases and confounding; and *iv*) the characterization of uncertainty for risk estimates. Other review papers about spatial analysis in environmental epidemiology exist (see *e.g.* Jerrett et al., 2010; Robertson et al., 2010; Carpenter, 2011; Meliker and Sloan, 2011; Osei, 2014), but most of them are devoted to single specific topics (*e.g.* clustering methods, surveillance methods, GIS) and keep a very general overview of the related statistical critical issues. On the other hand, the slant of our review paper is on the statistical aspects of data collection and analysis, and it provides a deeper insight into ecological spatial and spatio-temporal regression, while trying to be simple in the notation in order to be readable by a large community of scientists and researchers including geographers, epidemiologists, etc.

The rest of the paper is structured as follows. Section 2 introduces data commonly used in health and air pollution studies and discusses the most relevant issues, including exposure metric, spatial misalignment, preferential sampling, measurement error and ecological fallacy. Ecological spatial and spatio-temporal regression models are described in Section 3, with a detailed discussion on spatial random effects and multivariate models required in the case of multiple disease data. A short review of the available software for the considered models is given in Section 4, while discussion points are given in Section 5.

²<http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/>

2. Data in health and air pollution studies

Concerns about the health effects of air pollution from industrial, traffic and heating sources are currently the focus of considerable public and scientific attention. In general, the modelling approach used for assessing the impact of exposure on human health depends on the kind of data available and how the data are collected over space and/or time. Data used in studies to evaluate the exposure-risk relationship usually include disease events, measurements on pollutants or other risk factors and the reference population at risk, along with other covariates describing socio-demographic features. Air pollution concentrations are routinely measured at a number of monitoring sites across a region of interest producing point-reference data or are simulated by numerical models over a pre-set grid of locations. On the other hand, health data are mostly collected at an aggregated level and are usually associated with irregularly-shaped (administrative) regions. Epidemiological investigations in which associations between disease occurrence and environmental risk factors are studied over aggregated groups rather than on individual level are termed *ecological studies*. Even if these studies have been questioned because of their aggregated nature, they are particularly useful when individual level measurements of exposure are not available, as in case studies concerned with the investigation of the effect of air pollution on human health (Biggeri et al., 2004; Samet et al., 2000; Smith et al., 2000). This section provides a brief background on the kind of data available in air pollution health related studies and describes the potential problems of working with these types of data.

2.1. Health data

In statistical epidemiology, health data have usually been studied either in the form of count data or at individual level (case event data). In the latter case, two main types of studies can be distinguished: those based on analyzing the spatial pattern of case event locations using point processes (Grell et al., 2015; Diggle, 1990, Diggle, 2003, Pinto Junior et al., 2015), and cohort studies, which are useful to quantify the health effects resulting from long-term exposure (months or years) to pollution (see *e.g.* Molitor et al., 2006; Hoek et al., 2013). Firstly, they are not easily available and are costly to obtain due to the large amount of data required and the possible expenses related to GIS and global positioning system (GPS)-based devices (Nuckols et al., 2004). Secondly, the case home address could not be related to the disease aetiology. For instance, the disease could have been contracted in the working address because of occupational exposure or could be related to the time spent by individuals in several indoor and outdoor microenvironments characterized by different pollutant concentrations. In this regard some authors have proposed models for personal exposure based on longitudinal panel studies, time-activity diary data or GPS-enabled individual exposure monitor, see *e.g.* McBride et al. (2007), Zidek et al. (2007) and Blangiardo et al. (2011). As a result of these drawbacks related to individual data, either time series or areal unit study designs have also been used to quantify the short- and long-term health impact of air pollution. These studies utilise population level summaries rather than individual level data and cannot be used to quantify individual level cause and effect. However, both time series and areal unit (aggregated) data required to implement such studies are routinely collected by Health Authorities and are widely available. Although several studies have been proposed using both time series (Peng and Dominici, 2008) and spatial designs, here we only focus on areal data and the related spatial and spatio-temporal modelling challenges.

Health data are collected over time in a fixed study region, \mathcal{D}_y , typically in the form of mortality and morbidity counts or hospital admissions, coded according to the type of disease (*e.g.* cardiovascular, acute respiratory, etc). In general, occurrences of several, say n_y , diseases can be observed at a specific region (*e.g.* zip codes, counties), \mathbf{s}_i , and at a specific time point t . The complete set of information for the health data is thus denoted as

$$\{Y_k(\mathbf{s}_i, t), \mathbf{s}_i \in \mathcal{D}_y\}, \quad k = 1, \dots, n_y; i = 1, \dots, N_y; t = 1, \dots, T;$$

where k and i are the disease and region index, respectively. A great advantage of working with count data is their quick availability for a wide variety of diseases, particularly in countries with advanced statistical systems. This allows studying many variables and populations at relatively low cost (with respect to cases when data are not routinely collected). Because of its quick availability, this kind of data has been object of considerable methodological developments in the last years (see Section 3).

However, there are a number of problems than can arise from aggregating data into geographic units, including small number problem and frequent zero-valued observations (when events are counted in a short period of time and/or on very small areas for a rare disease; see Section 3.1.3), modifiable area unit problem (MAUP) and ecological

1
2
3
4 123 fallacy (see Section 2.3), and changes in geographic unit boundaries over time (Waller and Gotway, 2004; Meliker
5 124 and Sloan, 2011).

6 125 2.2. Air pollution data

7
8 Air pollution data are normally used as a measure of exposure in spatial and spatio-temporal regression models to
9 evaluate the associated health risk. Pollution concentrations are routinely measured at specific points in time and at
10 a number of monitoring sites across a continuous region \mathcal{D}_x and usually come in the form of geostatistical data. In
11 general, one or more pollutants can be observed at each monitoring site so that the entire set of information for these
12 data, at a specific monitoring site \mathbf{u}_l and a specific time point t , is denoted by

$$13 \{X_j(\mathbf{u}_l, t), \mathbf{u}_l \in \mathcal{D}_{x_j}\}, \quad j = 1, \dots, n_x; l = 1, \dots, N_{x_j}; t = 1, \dots, T;$$

14
15
16 126 where n_x is the number of observed pollutants while N_{x_j} is the number of monitoring sites for pollutant j . Note that this
17 127 notation includes the fact that pollutants can be measured by different monitoring networks, possibly characterized by
18 128 heterogeneity and/or heterotopicity, i.e. when pollutants are observed at non-collocated sites or with different sampling
19 129 strategies; see for example Fassò et al. (2007) and Fassò and Finazzi (2011). Additionally, pollution concentrations
20 130 estimated by computer dispersion models on a regular grid have also become freely available in recent times; when
21 131 these data are used, it turns out that the $\mathbf{u}_l, l = 1, \dots, N_{x_j}$, refer to the nodes, or interiors, of a regular rectangular lattice
22 132 \mathcal{D}_{x_j} .

23 133 Several studies have considered the long-term effects of air pollution on human health. Others have also tried to
24 134 estimate the health impact of short-term exposure to pollution (*i.e.* referring to a few days of elevated concentrations).
25 135 In practice, a number of statistical issues on how to best evaluate exposure to air pollution need to be addressed and
26 136 these may depend both on data availability and modelling strategy, as discussed in the following.

27 28 137 2.2.1. Exposure metric

29 138 An exposure metric is a way of summarizing a person's exposure to a particular element (*e.g.* air pollution) that can
30 139 then be used for hazard determinations. Ideally, one would use exposure estimates on an individual level, combining
31 140 air pollution data with relevant human activity patterns and mobility histories. This can be done using space-time
32 141 dynamics models in both environmental contaminants and GPS based mobility histories (see *e.g.* Elgethun et al.
33 142 2003; Berhane et al. 2004; Gerharz et al. 2009).

34 143 In practice, however, air pollution concentration measured at monitoring stations or predicted by a numerical
35 144 model is commonly used as a surrogate for individual exposure. Monitoring data may not be the most appropriate
36 145 exposure measure since it may not provide spatially representative pollution concentrations (Ozkaynak et al., 2013).
37 146 First, monitoring stations may have been put into place with a different aim, *e.g.* for regulatory purposes; in many
38 147 cases they are located close to important sources of pollution (see Section 2.2.3 on preferential sampling), resulting in
39 148 overestimated exposure (Meliker and Sloan, 2011; Shaddick and Zidek, 2014). Second, if the pollutants are spatially
40 149 homogeneous, it is reasonable to consider the monitored value as an average value for the whole area. Pollutants
41 150 associated to traffic, however, such as CO and NOx, tend to be spatially heterogeneous. In this case, the area-wide
42 151 exposure estimates can be obtained using spatio-temporal statistical models or computer-based models (see Section
43 152 2.2.2). The former are considered in a vast literature on air pollution modelling whose goal is spatial prediction at
44 153 unmonitored sites, that considers two predominant approaches: hierarchically structured models and geostatistical
45 154 models (see *e.g.* Zidek et al., 2002; Smith and Kolenikov, 2003; Sahu and Mardia, 2005; Sahu et al., 2006; Cocchi
46 155 et al., 2007; Fassò and Cameletti, 2009; Cameletti et al., 2011; Ignaccolo et al., 2014; Huang et al., 2015). The latter
47 156 simulate air pollutant transport, transformation and diffusion using input data on source emissions and meteorology
48 157 and provide numerical output in the form of gridded data (for example CMAQ³ or CHIMERE⁴).

49 158 Areal exposure estimates can be improved by integrating monitoring data in a statistical model with different data
50 159 sources such as remote sensing data (Fassò and Finazzi, 2011) or numerical model output (Van de Kassteele et al.,
51 160 2009; Bruno and Paci, 2013; Bruno et al., 2014; Ignaccolo et al., 2013; Paci et al., 2013; Huang et al., 2015). This is
52 161 usually referred to in the literature as *data fusion* (Berrocal et al., 2011; Paci et al., 2015).

53
54
55
56 ³<http://www.epa.gov/air-research/community-multi-scale-air-quality-cmaq-modeling-system-air-quality-management>

57 ⁴<http://www.lmd.polytechnique.fr/chimere/>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

162 Another choice to be made regarding the metric concerns the identification of suitable summary indicators (*e.g.*
163 maximum, average or other syntheses of pollution concentration). Huang et al. (2015) evaluated the impact of the
164 chosen summary on the estimated exposure effect (NO_2) on respiratory disease in Scotland over the period 2007-2011
165 and found that only maximum NO_2 (within each areal unit) appears to have a significant effect on the relative risk of
166 respiratory disease, while the area average does not.

167 In the literature there are also some works which take into account that air pollution is a heterogeneous mix of
168 several compounds of different size, source and toxicity. For example Atkinson et al. (2010) and Pirani et al. (2015)
169 consider PM_{10} and $\text{PM}_{2.5}$ concentration together with particle number concentration and composition (carbon, sulfate,
170 nitrate and chloride) for an epidemiological time series study. The limitation is that this kind of particle metrics are
171 usually collected at a limited number of monitoring sites and it is difficult to include a spatial dimension in the analysis.

172 2.2.2. *Spatial misalignment and Change of Support Problem (COSP)*

173 In this paper we refer to spatially misaligned data as data measured on different spatial supports (*i.e.* at point or
174 area level). For this reason, it is important to deliver statistical methods that are able to cope with different spatial
175 resolutions in order to effectively change the support of air pollution data (regressors) to achieve alignment with the
176 outcome measured at area level. Disease data appear as counts observed at irregularly-shaped regions $\mathbf{s}_i \in \mathcal{D}_y$. In
177 contrast, pollution data are typically available at points \mathbf{u}_l within a study region \mathcal{D}_{x_j} and appear irregularly spaced.
178 This spatial misalignment gives rise to the *Change of Support Problem (COSP)* which refers to the issues deriving
179 from making inference with data on different spatial supports (point or area level) and resolutions (*e.g.* zip, tract,
180 district, grid cell, pixel, etc.). As reviewed in Gotway and Young (2002) and Gelfand et al. (2010), COSP is a generic
181 term which encloses many cases as well as several solutions and approaches.

182 In spatial epidemiology, the COSP case consists in upscaling from point to area level. Usually, a two stage ap-
183 proach is adopted: the first stage predicts pollution exposure for each region \mathbf{s}_i ($i = 1, \dots, N_y$) using air pollution
184 concentrations $X_j(\mathbf{u}_l)$ available at the monitoring sites $\mathbf{u}_l, l = 1, \dots, N_{x_j}$; in the second stage these predictions are
185 linked with the health outcomes through a spatial regression model (see Section 3).

186 At the first stage a *naive approach* can be implemented and consists in computing the area-wide pollutant level as the
187 average of the values measured by the neighboring monitoring station(s), possibly using distance- or population-based
188 weights (Elliott et al., 2007; Madsen et al., 2008). This approach is reasonable for pollutants with a spatial homoge-
189 neous behavior but fails in representing the true concentration distribution especially when the number of monitors is
190 limited and they are not randomly placed throughout the considered area (Lee et al., 2015), or when particulate matter
191 constituents, which show a strong spatial heterogeneity, are taken into account (Krall et al., 2015). In these cases the
192 pollutant values measured at a single or limited number of monitors cannot be considered as an accurate estimate of
193 the average pollutant level across the area. When pollutant concentrations are available on a regular grid via numerical
194 air quality and atmospheric dispersion model, the realignment with the health data can be simply obtained by com-
195 puting an average over the intersections between the areas and the regular grid cells (see *e.g.* Rushworth et al., 2014;
196 Lee and Sarran, 2015).

197 An alternative consists in implementing a spatial model for the measured pollutant values in order to predict the
198 concentrations at some unobserved locations, and then average these predictions to obtain the exposure level over
199 the areas. In particular, point predictions can be obtained via geostatistical kriging, interpolation by inverse distance
200 weighting (*e.g.* Carlin et al., 1999; Hubbell et al., 2005; Bell, 2006; Young et al., 2009; Li et al., 2014), or via land
201 use regression (Hoek et al., 2008; Basagaña et al., 2013). A similar spatial prediction technique for point-block re-
202 alignment is block kriging (Cressie, 1993), which is a predictor of the average value over a region $\mathbf{s}_i \in \mathcal{D}_y$. Zhu et al.
203 (2003) present a Bayesian application of spatial upscaling based on block kriging implemented through Monte Carlo
204 approximation of integrals over grid points.

205 Even if easy and intuitive these approaches may not be able to provide an accurate estimate of the exposure distri-
206 bution. Recently a different line of research has been explored in several papers which consider the true pollution
207 level as a latent spatial (or spatio-temporal) continuous process, measured with error at a finite number of sites, and
208 defined through a stochastic model possibly including spatially varying covariates (*e.g.* meteorological variables) or
209 temporal dynamics. Furthermore, in a data fusion perspective, it is also possible to merge data coming from air qual-
210 ity numerical models or remote sensing observations. This approach is adopted for example by Wikle and Berliner
211 (2005), Fuentes and Raftery (2005), Fuentes et al. (2006), Choi et al. (2009), Peng and Bell (2010), Lawson et al.
212 (2012), Huang et al. (2015) and Blangiardo et al. (2016), with the possibility of obtaining the area level pollutant con-

centrations by averaging estimates available by sampling from the posterior predictive distribution at several locations chosen within an area on a given time point. All these aggregation and upscaling approaches provide area-wide exposure estimates which are usually plugged-in the linear predictor of the second stage health model (see Equation (1)) by assuming they are known covariates. This approach fails to properly consider the exposure uncertainty, possibly producing a biased estimation of air pollution risk on the health outcome. A first solution to this issue is the fully Bayesian model proposed by Lee and Shaddick (2010), which feeds the entire pollutant concentration posterior predictive distributions into the second stage, thus accounting for the uncertainty intrinsic in the concentration estimates. Alternatively, a measurement error approach can be adopted, like the one proposed for example by Gryparis et al. (2009) and Szpiro and Paciorek (2013) based on a Berkson error model which allows to reduce bias in the second stage estimation (health model). The Bayesian Maximum Entropy (BME) method is an alternative proposed by Lee et al. (2009) to deal with the COSP thus providing a framework for the non-linear integration of data obtained at different observation scales.

2.2.3. Preferential sampling

Typically, exposure is assessed using data from an existing monitoring network. However, air quality monitoring stations may be intentionally located in certain places; in many cases, they are located close to important sources of pollution, with the aim of monitoring compliance with air pollution regulation, resulting in overestimated exposure. When exposure is estimated at the area level, the assumption is that observed locations were selected at random, leading to inaccurate estimation of exposure (Shaddick and Zidek, 2014; Zidek et al., 2014) and hence, of its effect on health (Lee et al., 2015). Shaddick and Zidek (2014) show evidence of the effect of preferential sampling when assessing exposure. They found statistically significant differences in terms of pollution levels between stations within a monitoring network in the UK that were kept during the whole time period the network was working and stations that at some point were removed; *i.e.* stations at which pollution levels are low tend to be deleted from the network. Zidek et al. (2014) propose a general framework using a superpopulation modelling approach to adjust exposure estimates for preferential sampling. Lee et al. (2015) investigate the effect of preferential sampling on health effect estimates in a two-stage model via a simulation study and using real data; exposure is modelled at the first stage using data from a preferentially sampled network and from a randomly sampled network, leading to different conclusions about the effect of exposure on health, as well as having an effect on the variability and bias of the estimated effect.

2.2.4. Measurement error

To assess the effect of air pollution on the human health, accurate exposure measurements or estimates are necessary. Since it is usually not possible to measure individual exposure, a concentration field over an area at a certain time is considered as a surrogate. However this field can be measured only through a monitoring network providing measures at fixed locations with a certain time frequency. Thus a measurement error can occur, that can be due to instrument imprecision at a monitoring site as well as to prediction error of the model applied to obtain a concentration value at an unmonitored site. Moreover, when the spatial support of health data does not coincide with that of air pollutant data, the COSP arises and the misalignment leads to an error that can be taken into account in a proper model (see also Sections 2.2.2 and 3.1.1). Also, concentration fields may be often obtained as output of numerical models that are affected by uncertainty related to initial conditions, parameters in model equations as well as model structure (Bayarri et al., 2009; Paci et al., 2015); such uncertainty constitutes a calibration error that could propagate to the response when model output is used as a predictor. To take into account this error and combine numerical output with observations, the *data fusion* and *data assimilation* techniques have been developed (Gelfand et al., 2010; Evensen, 2009). Data fusion consists in merging data coming from various sources of different nature and in the spatial context it requests to deal with COSP (see Section 2.2.2). Instead, data assimilation, which can be seen as a particular case of data fusion, deals with the inclusion of observed data into numerical models to estimate initial conditions in a predictive model consistent with the available observations and the underlying model dynamics.

In general, to include measurement errors the so-called *error-in-variables* (EIV) models are considered, often referred to as *measurement error* models (MEM); for a general overview see Fuller (1987) for linear models and Carroll et al. (2006) for nonlinear ones. Two types of MEM are distinguishable: the classical error model and the Berkson error model (Shaddick and Zidek, 2015). Under the first one, the measurement model equation states that the observed exposure is randomly distributed around the true latent level in the additive version (see *e.g.* Van de Kasstele and Stein, 2006). Instead the Berkson error model specifies the distribution of the true unobserved exposure

as being dependent on the observed exposure measurement. Both models can be additive or multiplicative, and a spatial component can be incorporated in the model error term (Gray et al., 2011). For spatial exposure, some works (Gryparis et al., 2009; Szpiro et al., 2011; Sheppard et al., 2012; Szpiro and Paciorek, 2013) decompose the measurement error into a Berkson-like component from smoothing the exposure surface and a classical-like component from variability in estimating exposure model parameters, even considering spatial misaligned data (see also Section 2.2.2).

2.3. Ecological Fallacy and MAUP

Inferences about individuals at locations are functions of regional average characteristics that are ascribed to a particular point in space and to a particular individual; assuming that associations observed at area level hold for the individuals within the areas can lead to the so-called ecological fallacy (Piantadosi et al., 1988; Greenland and Morgenstern, 1989; Greenland and Robins, 1994; Clayton et al., 1993; Morgenstern, 1998; Wakefield and Elliot, 1999; Wakefield, 2003). Several works have investigated the implications of ecological fallacy and the inherent bias under a modelling perspective (see Wakefield and Lyons (2010) and references therein). Given a non-linear model for the individual risks, such model changes when individual risks are aggregated at the areal level (specification bias). Wakefield and Shaddick (2006) use this modelling framework to show that specification bias is small if the within area variability in exposure is close to zero. They also point out that “the key to minimizing ecological bias is to have a fine enough partition of space at which exposure measurements are available, relative to the spatial exposure variability”. As discussed by Wakefield and Lyons (2010), the only way of accounting for the loss of information implied by working with areal data (and therefore alleviate the ecological fallacy problem) is to include individual level data in the modelling process. In particular, Jackson et al. (2008) proposed what they have termed as *hierarchical related regression* to jointly model aggregated and individual data, improving ecological inference.

Inference based on count data is subject to bias due to their aggregate nature. This is a problem found in non-spatial epidemiological studies (see for example Breslow and Day, 1980) and is amplified in spatial studies, where the effect of aggregation on spatial correlation constitutes an additional source of bias. From the spatial point of view, ecological fallacy arises because inference may change depending on how data are aggregated into geographical regions. The spatial manifestation of ecological fallacy has been recognised as a particular case of the Modifiable Areal Unit Problem (MAUP, Gotway and Young, 2002), which comprises two interrelated problems: the scale effect and the zoning effect (Openshaw and Taylor, 1981). In the former, inference changes when data are grouped into increasingly large geographical regions; this is the analogous of aggregation bias in non-spatial ecological studies arising when individuals are grouped. The source of this problem is the smoothing effect resulting from averaging, which reduces heterogeneity among units, or equivalently, some information about the spatial variability of the cases is lost. Aggregation-induced bias depends on the heterogeneity in the grouped observations: a completely homogeneous grouping system would be free from this problem (Openshaw, 1984). Indeed, as first observed in the pioneering work of Roberston (1950), the correlation between two variables measured at an ecological level can be expressed as the sum of the within group and between group components; when the within group component is negligible the ecological bias vanishes. This means that the smoothing effect is alleviated if the original observations are characterised by positive spatial correlation. On the other hand, the zoning effect arises when alternative formations of the areal units lead to different results; if the variation among areal units is not constant over the study region, different zoning rules can lead to different spatial correlation structures. This is the analogous of the grouping effect in non-spatial ecological studies.

3. Spatial and spatio-temporal regression

The association between air pollution and health data has been typically studied through the development of three-level hierarchical regression models. Such models have an intuitive appeal and enjoy several advantages. For example, they are well-suited for incorporating the foregoing knowledge at various levels of the modelling, are easy to interpret and facilitate model fitting. Furthermore, within the Bayesian paradigm, they enable exact inference and proper uncertainty assessment within the given specification.

In general, at the first level, the health data are conditioned on the process and parameters. The data can be conditioned on whatever aspects of the process are appropriate. At the second level of the hierarchy the process

component is specified. This component can have multiple levels, it can be spatial and/or dynamic and its stochastic form can be univariate or multivariate. Finally, at the third level, hyperprior distributions are specified (Gelfand, 2012).

In this section, we start by critically reviewing procedures for prior choices used in ecological spatial regression models for count data and then extend the study to spatio-temporal regression. These models are typically developed for modelling single diseases. However, in many cases, joint modelling of diseases can increase the understanding of diseases dynamics and relationships among diseases incidence. The merit of joint modelling can be high if the considered diseases share environmental risk factors. This has led to the development of multivariate spatial methods suited for areal data. The last part of the Section is thus dedicated to the discussion of multivariate spatial regression and the specification of spatial and spatially dynamic factor models, followed by a brief discussion on spatio-temporal regression models.

3.1. Univariate spatial regression

Suppose that the region of interest, \mathcal{D}_y , is split into N_y contiguous areas and that the observed number of disease cases in an area is denoted as $Y(\mathbf{s}_i)$, with $\mathbf{s}_i \in \mathcal{D}_y$. To assess which areas exhibit elevated or low levels of disease risk, the number of cases expected to occur in each area, $E(\mathbf{s}_i)$, is calculated; expected counts are thought as fixed and known functions of the size and demographic structure of the population living within each area (Banerjee et al., 2004; Lawson, 2009).

The number of expected cases (calculated using internal standardization) is what would be observed if the disease risk was constant over the whole study region and the spatial variations in incidence were due uniquely to population density and structure. In disease mapping studies, the focus is on identifying features of the spatial distribution of the disease rate that depart from what expected under the assumption of a constant disease risk.

If $E(\mathbf{s}_i)$ is not too large (*i.e.* the disease is rare and/or the population at risk is small), a standard univariate spatial model for the $Y(\mathbf{s}_i)$ is given by the Poisson model (Lawson, 2009)

$$Y(\mathbf{s}_i) | \eta(\mathbf{s}_i) \stackrel{ind}{\sim} Po(E(\mathbf{s}_i) e^{\eta(\mathbf{s}_i)})$$

with $\eta(\mathbf{s}_i)$ being the log relative risk in areal unit \mathbf{s}_i . This model completes with the linear predictor

$$\eta(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \phi(\mathbf{s}_i) \quad (1)$$

where $\mu(\mathbf{s}_i)$ is a mean level component and $\phi(\mathbf{s}_i)$ is a random effect introduced to capture any residual spatial autocorrelation present in the data.

3.1.1. Modelling the mean component

The simplest specification of the mean component appears as a parameterised function of the type, $\mu(\mathbf{s}_i) = \bar{\mathbf{x}}(\mathbf{s}_i)' \boldsymbol{\beta}$. The associated vector of regression coefficients, $\boldsymbol{\beta}$, is of main interest and helps in evaluating the impact of exposure variables on the health outcome under examination. Because several sources of uncertainty are in place, several statistical issues must be considered when trying to reliably estimate $\boldsymbol{\beta}$.

Vector $\bar{\mathbf{x}}(\mathbf{s}_i)$ denotes exposure variables at area \mathbf{s}_i . Usually, exposure variables are not directly observed at area level and the pollution concentrations must be estimated on administrative regions by applying COSP methods as discussed in Section 2.2.2. Also, the Poisson log-linear model considers the vector of estimated pollution concentrations as known and assumed to be constant across each areal unit. The hierarchical modelling approach allows to address the misalignment between health and exposure data in a straightforward manner, by simply adding a further level for Bayesian spatial prediction to estimate the pollution concentrations for all the areal units. The specification of a measurement error model for the exposure also allows to treat such exposure variables as model parameters and fully Bayesian inference guarantees that uncertainty concerning the parameters defined at lower levels correctly propagates through the hierarchy, giving proper assessment of the uncertainty for $\boldsymbol{\beta}$ at the first level.

There are also two major difficulties in making inference on regression parameters. Firstly, exposure variables in $\bar{\mathbf{x}}(\mathbf{s}_i)$ might be themselves highly correlated, leading to multicollinearity. There exist several broad strategies to deal with multicollinearity and typical examples are represented by ridge regression and partial least squares regression (Brown, 1993). Another possibility is to perform a dimensionality reduction procedure on exposure variables by means of principal components or factor analysis; the latter is discussed in Section 3.2.2.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

353 A second issue regards allowance for the effect of confounders into the model. A variable is a confounding factor
354 for the effect of air pollution on health when it is correlated with both the health outcome and the associated exposure
355 variables. For a formal definition of confounding in epidemiological studies and possible solutions to deal with it see,
356 for example, Rothman et al. (2012). Typical confounders considered in the literature include age, gender, ethnicity
357 and socio-economic deprivation. Meteorological variables, such as temperature and humidity, are also well known
358 to be confounders when analysing the relationship between air pollution and mortality, especially in short-term effect
359 studies. Not accounting for all potential confounders may cause severe bias in estimating β and can lead to wrong
360 conclusions on the effect of exposure on health (Shaddick and Zidek, 2015).

361 It is important to distinguish between unmeasured and measured confounders (Peng et al., 2006). When con-
362 founding variables are measured, standard practice is to include them into the mean component in model (1), so that
363 $\mu(\mathbf{s}_i) = \tilde{\mathbf{x}}(\mathbf{s}_i)' \beta + \tilde{\mathbf{z}}_m(\mathbf{s}_i)' \beta_z$, where $\tilde{\mathbf{z}}_m(\mathbf{s}_i)$ is the vector of measured confounders at areal unit \mathbf{s}_i and β_z is the associated
364 vector of parameters.

365 Managing the effect of unmeasured confounders is a more controversial task to be addressed. A popular approach
366 is to assume that these unmeasured confounders drive the residual spatial correlation, after accounting for exposure
367 and measured confounders in the mean component, and that this variability can be captured by the spatial random
368 effects, $\phi(\mathbf{s}_i)$, in model (1). Hence, omitting the random effects, $\phi(\mathbf{s}_i)$, from the model is not a sensible choice as its
369 dependence with the covariates causes bias in estimating β .

370 3.1.2. Modelling the spatial random effect

371 Since the random variables $\phi(\mathbf{s}_i)$ in equation (1) are devoted to capture area-level spatial effects that are represen-
372 tative of a whole region, models suited for lattice data, *i.e.* Gaussian Markov Random Fields (GMRF, see Rue and
373 Held, 2005 for a comprehensive overview), arise as a natural choice in this context and are almost the standard in
374 the literature concerning ecological spatial regression. For this reason, they receive major attention in this Section.
375 Alternatively, an example of a geostatistical approach is given in Kelsall and Wakefield (2002), where the correlation
376 structure is derived through consideration of an underlying continuous risk surface. The multivariate Normal distribu-
377 tion is one of the most flexible distributions for representing spatially correlated random variables. For instance,
378 writing $\phi = (\phi(\mathbf{s}_1), \dots, \phi(\mathbf{s}_{N_y}))$, one might assume that $\phi | \varphi \sim N(\mathbf{0}, \Sigma(\varphi))$, where $\Sigma(\varphi)_{ij}$ gives the covariance between
379 $\phi(\mathbf{s}_i)$ and $\phi(\mathbf{s}_j)$ as a function of some hyperparameters φ (Banerjee et al., 2004).

380 However, in health care research, there is rarely much substantive knowledge to guide the choice of the covari-
381 ance function, and often quite weak information in the data to estimate the parameters of this function, particularly
382 for more complex forms. As high long range correlation of the risks is difficult to distinguish from the effect of the
383 overall mean, it is also important to ensure that the chosen correlation function (and associated hyperpriors) gives near
384 zero correlation at distances within the extent of the study region, to avoid nonidentifiability of the mean and corre-
385 lation parameters (Best et al., 2005). Moreover, a further important practical limitation of this approach is that with
386 large study regions the implementation via Markov chain Monte Carlo (MCMC) algorithms can be computationally
387 expensive due to the large amount of matrix inversion required at each iteration.

As stated before, the most relevant class of models in the context of regression involving an area-level response is
that of GMRFs, also known as conditional autoregressive (CAR) models (see, *e.g.*, Cressie, 1993, sec. 6.3.2). These
models specify conditional dependence involving a (usually small) set of spatial neighbours. The specification of
the neighbourhood structure is commonly based on the adjacency matrix \mathbf{W} , which is a symmetric 0/1 matrix with
element w_{ij} equal to 1 if i and j are neighbours (by definition, $w_{ii} = 0$). CAR models are specified starting by N_y
conditional distributions:

$$E(\phi(\mathbf{s}_i) | \phi(-\mathbf{s}_i)) = \sum_j \alpha_{ij} \phi(\mathbf{s}_j), \quad \text{Var}(\phi(\mathbf{s}_i) | \phi(-\mathbf{s}_i)) = \kappa(\mathbf{s}_i),$$

388 where $\phi(-\mathbf{s}_i)$ denotes all other values but $\phi(\mathbf{s}_i)$, $\alpha_{ij} = \rho w_{ij} / w_{i+}$, with $w_{i+} = \sum_j w_{ij}$, and $\kappa(\mathbf{s}_i) = \tau^2 / w_{i+}$. The α_{ij} are
389 coefficients reflecting local spatial dependence between units i and j while $\kappa(\mathbf{s}_i)$ is the conditional variance which is
390 inversely proportional to the number of neighbours. Note that this specification for α_{ij} and $\kappa(\mathbf{s}_i)$ satisfies the symmetry
391 condition, $\alpha_{ij} \kappa(\mathbf{s}_j) = \alpha_{ji} \kappa(\mathbf{s}_i)$, which is a necessary condition for the joint distribution of ϕ to be valid (Cressie, 1993).

392 The parameter ρ can be thought of as an autocorrelation parameter that reflects the overall strength of spatial
393 dependence between locations with nonzero weights. To define a proper joint distribution, the covariance (or the

precision) matrix must also be positive definite: this requires that $|\rho| < 1$ if the scaled adjacency matrix, $D^{-1}W$, where $D = \text{diag}(w_{i+})$, is used (Cressie, 1993).

An appealing feature of the GMRF prior is the possibility to make inference about the overall degree of spatial dependence in the disease risk by estimating ρ . However, interpretation of ρ is not straightforward and values close to the maximum are needed to reflect even moderate spatial dependence. This drawback led Besag et al. (1991) to consider the intrinsic CAR model as a prior for ϕ . This model is obtained by setting ρ to its upper limiting value of 1 such that the conditional expectations for $\phi(s_i)$ are equal to the mean of the random effects in neighbouring areas. This is the simplest possible CAR prior and is appropriate if the residuals from the covariate component of the linear predictor are spatially smooth across the entire region. Also, although the univariate conditional prior distributions are well defined, the corresponding joint multivariate Gaussian distribution of the intrinsic CAR is improper since the precision matrix is singular. Finally, notice that the improper CAR prior is a pairwise difference prior that can be identified only up to an additive constant. Hence, to identify an intercept term in the linear predictor, a sum-to-zero constraint on the random effects is needed (Banerjee et al., 2004).

The intrinsic CAR model is suited for capturing unexplained spatially structured variability. In order to capture both spatially structured and unstructured variability, Besag et al. (1991) proposed to introduce a second set of independent Gaussian random effects, say $\nu(s_i)$, with mean zero and a common variance. Following this model specification, which is also known as the *Besag-York-Mollié* (BYM) or convolution model, different levels of spatial smoothness can be achieved by varying the relative sizes of the two components $\phi(s_i)$ and $\nu(s_i)$. However, the disadvantage of this flexibility is that each data point is represented by two random effects, and hence only their sum is identifiable. A thorough discussion about difficulties concerning model estimation and identifiability of area-specific random effects $\phi(s_i)$ and $\nu(s_i)$ is provided in Eberly and Carlin (2000).

Alternatives to the convolution model, which avoid the potential identifiability problem encountered with the BYM prior, have been proposed by MacNab (2003) and Stern and Cressie (2000). However, these extensions consider a global smoothing parameter, such that the amount of smoothing performed is affected globally by all the areas and is not adaptive. This could be inappropriate for two reasons. Firstly, since the spatial distribution of air pollutant is smooth, the residual spatial structure obtained after considering the covariate effect is unlikely to be globally spatially smooth, and is instead likely to exhibit localised smoothness (Lee et al., 2014). Secondly, if the spatial random effects capture the effect of spatially correlated unmeasured variables, potential collinearity between this component and any other spatially smooth covariate can lead to variance inflation and bias in the estimation of the air pollution effect, essentially because of spatial confounding (Hodges and Reich, 2010; Paciorek, 2010). As discussed by Paciorek (2010), the size of the spatial confounding bias depends on the spatial scale of the variability in the outcome and the exposure and can only be reduced if the unmeasured confounders (in the form of spatial random effects) are responsible for the large-scale spatial variability in the outcome, with exposure only explaining the small-scale spatial patterns.

To our knowledge, approaches to solving these problems have been proposed either by spatially constrained regression, where the spatial random effects are orthogonal to the covariates (see, for example, Reich et al., 2006, Hodges and Reich, 2010, and Hughes and Haran, 2013), or relaxing the global smoothing restrictions of the CAR prior to allow for *localized* priors as discussed in Lee et al. (2014). Other authors have also developed semiparametric spatial models, that replace the continuously varying spatial distribution for the log-relative risk by discrete allocation or partition models with each cluster or component having a constant unknown relative risk. A critical discussion of this approach can be found in Best et al. (2005).

3.1.3. Departures from the Poisson model

The spatial epidemiological literature often refers to extra-Poisson variability in the data to indicate any departure from the Poisson model (i.e. variance equal to the mean). The most frequent situation is overdispersion, which is when the health outcome has variance larger than the mean. The introduction of the spatial random effects $\phi(s_i)$ in (1) can be seen as a device for inducing over-dispersion, or extra-Poisson variation, into the model. More precisely, $\phi(s_i)$ captures extra-Poisson variation due to spatial correlation (which may be induced by unobserved confounders varying at the areal level, e.g. socio-economic factors; see also discussion in Section 3.1.1). Extra-Poisson variability can also be driven by spatially unstructured unobserved confounders, i.e. varying within areas at an individual level (e.g. smoking), or simply be caused by anomalies in the data (Wakefield and Elliot, 1999). In general, models which fail to account for overdispersion lead to underestimation of the uncertainty associated to the estimates of the log

relative risks $\eta(\mathbf{s}_i)$ (Shaddick and Zidek, 2015). A popular model for overdispersed counts is the negative binomial distribution, i.e. a Poisson with a Gamma distribution on the mean parameter, which offers a suitable approach to deal with spatially unstructured extra-Poisson variability in the data. Based on a Poisson-Gamma framework, Wakefield (2007) investigates several parametrizations for the Gamma distribution giving different variance mean relationships. The framework for regression presented throughout Section 3 allows to accommodate spatial extra-Poisson variability by means of area-level random effects in a Poisson log-Normal model. A more general approach, which is able to cope both with under and overdispersion, consists in using a Generalised Poisson Likelihood, see Fuentes et al. (2006) for an application of this model in evaluating the effect of particulate matter pollution on health.

When dealing with a data set with an excessive number of zeros, zero-inflated models can be used. The zero inflated Poisson (Agarwal et al., 2002) model has received a lot of attention and has always been used in modelling count data where the extra variations are solely caused by the extra zeros. For health data where the extra variability is caused by excess zeros and also unobserved heterogeneity, recommended models may be, for example, zero inflated negative binomial and zero inflated generalized Poisson (Gschlößl and Czado, 2008).

3.2. Multivariate extensions

Many health care research studies refer to the joint study of multiple diseases. In general, the abundance of measures is both an opportunity and a challenge from the epidemiological and the statistical point of view. From the epidemiological point of view, joint modelling of diseases can increase the understanding of diseases dynamics and relationships among diseases occurrence. The merit of joint modelling can be high if the considered diseases share risk factors or if the presence of a disease encourages (or inhibits) the occurrence of another one. From the statistical point of view, an evident advantage in joint disease mapping is that, if the disease risks are correlated, standard errors of the estimates obtained via univariate modelling can be sensibly reduced. Moreover, estimates for rare diseases can borrow strength from more diffuse diseases. In fact, correlation between diseases within areas, between areas within diseases and between areas and diseases constitutes valuable information contained in the data that can be used to increase efficiency of the estimates.

Modelling several diseases jointly in an effective way, thereby borrowing information across them, is a difficult task in general. While models for univariate disease data have been extensively explored, models for multivariate lattice data have only been developed in the past few years.

Let $Y_k(\mathbf{s}_i)$ and $E_k(\mathbf{s}_i)$ denote the observed and expected number of cases for disease $k = 1, \dots, n_y$ in a region \mathbf{s}_i , with $\mathbf{s}_i \in \mathcal{D}_y$. As in Section 3.1, we consider the case where health data are modelled as (conditionally) independent Poisson variables, so that

$$Y_k(\mathbf{s}_i) | \eta_k(\mathbf{s}_i) \stackrel{ind}{\sim} Po(E_k(\mathbf{s}_i) e^{\eta_k(\mathbf{s}_i)})$$

with $\eta_k(\mathbf{s}_i)$ being the log relative risk in areal unit \mathbf{s}_i for disease k . The linear predictor can be specified as

$$\eta_k(\mathbf{s}_i) = \mu_k(\mathbf{s}_i) + \phi_k(\mathbf{s}_i) \quad (2)$$

where $\mu_k(\mathbf{s}_i)$ and $\phi_k(\mathbf{s}_i)$ represent the mean level component and the spatial random effect at area \mathbf{s}_i for disease k , respectively. In matrix form, the linear predictor in equation (2) can be written as

$$\boldsymbol{\eta}(\mathbf{s}_i) = \boldsymbol{\mu}(\mathbf{s}_i) + \boldsymbol{\phi}(\mathbf{s}_i), \quad (3)$$

where $\boldsymbol{\mu}(\mathbf{s}_i)$ and $\boldsymbol{\phi}(\mathbf{s}_i)$ are n_y -dimensional vectors of fixed and random effects, respectively, for the n_y diseases at site \mathbf{s}_i . In general, multivariate models differ mainly with respect to the specification of the joint distribution of the random effects and, more specifically, on how dependences within and between diseases are defined. Henceforth, we shall discuss solutions which either generalise CAR models to a multivariate framework or follow a factor-analytic approach.

3.2.1. Multivariate CAR models

Univariate CAR models are defined by the conditional distribution of the variable at each site given the variable at all the other sites. The extension to the multivariate case follows the same rationale. The more usual generalisation (Mardia, 1988), considers the N_y Gaussian conditional distributions $\boldsymbol{\phi}(\mathbf{s}_i) | \boldsymbol{\phi}(-\mathbf{s}_i)$ of the n_y -vector variable $\boldsymbol{\phi}(\mathbf{s}_i)$ at each

site s_i given $\boldsymbol{\phi}(-s_i)$, where $\boldsymbol{\phi}(-s_i)$ denotes the values $\boldsymbol{\phi}(s_j)$ at all sites $s_j \neq s_i$. The conditional expectations and variances are specified as

$$E(\boldsymbol{\phi}(s_i)|\boldsymbol{\phi}(-s_i)) = \sum_j \mathbf{A}_{ij}\boldsymbol{\phi}(s_j), \quad \text{Var}(\boldsymbol{\phi}(s_i)|\boldsymbol{\phi}(-s_i)) = \boldsymbol{\Gamma}_i,$$

where \mathbf{A}_{ij} are matrices of parameters relating to the cross-dependences of the variables and $\boldsymbol{\Gamma}_i$ is the $n_y \times n_y$ conditional covariance matrix. Again, as in the univariate case, the conditional dependence structure is limited to neighboring areas identified by the adjacency matrix \mathbf{W} such that $\mathbf{A}_{ij} \neq \mathbf{0}$ if and only if $w_{ij} \neq 0$.

Let $\boldsymbol{\phi}_s = (\boldsymbol{\phi}(s_1)', \dots, \boldsymbol{\phi}(s_{N_y})')'$ denote the $N_y n_y$ -vector of random effects ordered by site. Also, let us define $\boldsymbol{\phi}_v = (\boldsymbol{\phi}_1(s_1, \dots, s_{N_y})', \dots, \boldsymbol{\phi}_{n_y}(s_1, \dots, s_{N_y})')'$ as the same vector with random effects ordered by variables, which shows clearly the relevant contribution of each disease and the joint contributions of pairs of diseases. Clearly, the two forms are just different ways of representing the same multivariate process. One form may sometimes be more convenient to specify than the other. Under Gaussianity, given the N_y conditional distributions, the joint distribution of $\boldsymbol{\phi}_s$ is multivariate Gaussian with mean zero and precision matrix $\mathbf{P}_s = \text{Block}(-\boldsymbol{\Gamma}_i^{-1}\mathbf{A}_{ij})$; the precision matrix \mathbf{P}_v can be obtained by permuting the rows and the columns of \mathbf{P}_s . In order to satisfy the symmetry condition for the joint covariance matrix, the equality $\mathbf{A}_{ij}\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}_i\mathbf{A}'_{ji}$ needs to hold. One difficulty concerning this approach is that conditions on the smoothing parameters for positive definiteness of the joint distribution precision matrix depend on the unknown conditional covariance matrix $\boldsymbol{\Gamma}$. A general strategy for checking positive definiteness of the precision matrices, as derived below, is proposed in Ippoliti et al. (2015).

Assumptions are needed to reduce the total number of parameters. For example, it is often reasonable to assume that the conditional variance $\boldsymbol{\Gamma}_i$ is either constant, or only varies over sites by a constant, so that $\boldsymbol{\Gamma}_i = w_{i+}^{-1}\boldsymbol{\Gamma}$. The multivariate generalisations of univariate CAR models proposed in the literature, mainly differ in the way in which both the spatial correlation and the cross-correlations are managed. A popular and fairly general approach is proposed by Jin et al. (2007). The approach is based on the linear model of coregionalization and encompasses several other models as special cases. The specification of the model begins with the definition of latent spatial effects, $\boldsymbol{\omega}_v = (\boldsymbol{\omega}'_1, \dots, \boldsymbol{\omega}'_k, \dots, \boldsymbol{\omega}'_{n_y})'$, where $\boldsymbol{\omega}_k = (\omega_k(s_1), \dots, \omega_k(s_{N_y}))$, $k = 1, \dots, n_y$, is a N_y -dimensional spatial process. The spatial effects for each disease are then modelled as $\boldsymbol{\phi}_v = (\mathbf{G} \otimes \mathbf{I}_{N_y})\boldsymbol{\omega}_v$, where \mathbf{G} is the upper triangular Cholesky decomposition of $\boldsymbol{\Gamma}$. A versatile model is obtained by assuming that the $\boldsymbol{\omega}_k$ are dependent and not identical latent processes so that the joint distribution of $\boldsymbol{\omega}_v$ is $MVN(\mathbf{0}, (\mathbf{I}_{n_y} \otimes \mathbf{D} - \mathbf{B} \otimes \mathbf{W})^{-1})$, where \mathbf{B} is a n_y -dimensional symmetric matrix whose diagonal entries capture the spatial autocorrelation for the k -th disease, and off-diagonal entries capture the spatial cross-correlation between diseases k and k' . This model is dubbed $MCAR(\mathbf{B}, \boldsymbol{\Gamma})$ in what follows. It turns out that the joint distribution of the vector $\boldsymbol{\phi}_v$ is a zero mean multivariate Gaussian with precision matrix

$$\mathbf{P}_v = (\mathbf{G} \otimes \mathbf{I}_{N_y})(\mathbf{I}_{n_y} \otimes \mathbf{D} - \mathbf{B} \otimes \mathbf{W})^{-1}(\mathbf{G} \otimes \mathbf{I}_{N_y})'.$$

Conditions for positive definiteness of the joint precision matrix are obtained by constraining the eigenvalues of matrix \mathbf{B} to lie between the reciprocal of the minimum and maximum eigenvalues of $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. The $MCAR(\mathbf{B}, \boldsymbol{\Gamma})$ model belongs to the class of multivariate CAR models that explicitly model symmetric cross-covariances, and allows different disease-specific smoothing parameters.

A less general model (Carlin and Banerjee, 2003), which turns out to be a special case of the $MCAR(\mathbf{B}, \boldsymbol{\Gamma})$ model, assumes one smoothing parameter for each disease and can be obtained by setting $\mathbf{B} = \text{diag}(\rho_1, \dots, \rho_{n_y})$, delivering the $MCAR(\rho_1, \dots, \rho_{n_y}, \boldsymbol{\Gamma})$ model. Conditions for positive definiteness require that $|\rho_k| < 1$, $k = 1, \dots, n_y$.

The $MCAR$ model proposed in Gelfand and Vounatsou (2003) assumes a common spatial smoothing parameter ρ capturing disease-specific correlation and cross-correlations, such that the joint precision matrix can be obtained by setting $\mathbf{B} = \rho\mathbf{I}_{n_y}$, delivering the $MCAR(\rho, \boldsymbol{\Gamma})$. According to this specification, the condition for positive definiteness of the covariance matrix reduces to $|\rho| < 1$ as in the univariate framework. This approach delivers an immediate and intuitive generalisation to the multivariate framework with the evident drawback that a single smoothing parameter ρ is employed to capture disease-specific spatial association and cross covariances. Note that the multivariate intrinsic CAR distribution is obtained if $\rho = 1$.

The $MCAR(\rho, \boldsymbol{\Gamma})$, $MCAR(\rho_1, \dots, \rho_{n_y}, \boldsymbol{\Gamma})$ and $MCAR(\mathbf{B}, \boldsymbol{\Gamma})$ models improve each other with respect to the generality of the spatial correlation and cross-correlation structure considered, the latter model being the most general one, with the only unsatisfactory restriction of specifying symmetric cross-covariances. A first generalisation of the CAR distribution to a multivariate framework allowing non-symmetric cross-covariances was proposed by Kim et al.

(2001), by specifying *univariate* full conditionals; however the extension of the model to the case where $n_y > 2$ appears unfeasible.

In Jin et al. (2005) an approach based on the direct specification of the joint distribution for the multivariate spatial process through the specification of marginal and conditional distributions is proposed. Although this approach allows for the specification of asymmetric covariance structures, it has the main drawback that the conditioning order has an effect on the results and it is difficult to apply when n_y is large.

A more general $MCAR(\tilde{\mathbf{B}}, \mathbf{\Gamma})$ asymmetric model, suited for the case where $n_y > 2$, has been proposed by Greco and Trivisano (2009). As for the $MCAR(\mathbf{B}, \mathbf{\Gamma})$ model, the model of Greco and Trivisano (2009) starts from the linear model of coregionalisation but allows the off-diagonal elements of \mathbf{B} to be different, obtaining the joint covariance matrix as

$$P_V = (\mathbf{G} \otimes \mathbf{I}_{N_y})(\mathbf{I}_{n_y} \otimes \mathbf{D} - (\tilde{\mathbf{B}} \otimes \mathbf{W}_U + \tilde{\mathbf{B}}' \otimes \mathbf{W}'_U))^{-1}(\mathbf{G} \otimes \mathbf{I}_{N_y})'$$

where \mathbf{W}_U denotes the lower triangular part of the adjacency matrix \mathbf{W} and $\tilde{\mathbf{B}}_{kk'} \neq \tilde{\mathbf{B}}_{k'k}$. Conditions for positive definiteness of the joint covariance matrix are obtained by constraining the singular values of $\tilde{\mathbf{B}}$ to lie between 0 and 1. The $MCAR(\mathbf{B}, \mathbf{\Gamma})$ model arises as a special case of this model when $\tilde{\mathbf{B}}$ is constrained to be symmetric.

The model proposed by Sain and Cressie (2007), dubbed by the authors as the canonical multivariate conditional autoregressive model (CAMCAR), is characterized by a very similar specification with more restrictive conditions for positive definiteness with respect to those found in Greco and Trivisano (2009). Sain et al. (2011) also provide a different representation of multivariate lattice data that was suggested, but not implemented, in Sain and Cressie (2007). The idea is to think of the multivariate process as a univariate CAR over $N_y n_y$ regions, so that the process can be expressed through the individual conditional distributions, $\phi_k(\mathbf{s}_i) | \phi_{-k}(-\mathbf{s}_i)$, of each variable at each site. This formulation of a multivariate MRF allows great flexibility in modelling the conditional dependence structure and is easily extendable to complex neighborhood structures.

It is also worth mentioning that a partial review on GMRF/CAR and multivariate GMRF prior formulations, both for univariate and multivariate disease mapping models, is presented in MacNab (2011) giving insights into various prior characteristics for representing disease risks variability and spatial interaction.

Finally we note that, together with multivariate CARs, purely factor models (see Section 3.2.2) and smoothed analysis of variance (Zhang et al., 2009) also represent an alternative approach for the analysis of multiple diseases. Martinez-Beneito (2013) describes a general modelling approach that combines several different spatial structures with different multivariate dependence schemes. An alternative reformulation that accrues substantial computational benefits enabling the joint mapping of several diseases is also discussed by Botella-Rocamora et al. (2015).

3.2.2. Latent variable models

Multivariate spatial analysis becomes computationally onerous as soon as the number of geographical units is large and more than two or three diseases are available - see, for example, Dobra et al. (2011). Therefore, the development of computationally efficient multivariate models is desirable.

One main question, both from a methodological and an applied standpoint, is how to condense the available information into interpretable aggregates. One possibility is to use Factor Analysis (FA), which is a dimension reduction method to search for the underlying structures of multiple diseases. In practice, it is assumed that diseases that are from the same population groups or areas, often exhibit similar characteristics leading to the belief that they might be driven by some common sources, often referred to as common factors.

The use of a factor-analytic approach enjoys several advantages. For example, with the identification of a (usually small) number of latent variables, FA avoids the curse of dimensionality commonly present in large spatial data. Also, since the variables are often correlated with each other and each of the variables might also be correlated across the locations due to geographic similarities, FA facilitates the identification of spatial clusters in the latent factors which further avoids dimensionality issues. Finally, for studies considering the combined effects of multiple pollutants simultaneously, there might be the need of constructing appropriate air quality indicators based on dimensionality reduction techniques (Rushworth et al., 2014). In this case, FA enables all available pollutants to be combined and used to estimate the health impact of a proxy measure of the air pollution in ecological-type regression.

A spatial factor model can be specified through the linear predictor

$$\eta_k(\mathbf{s}_i) = \mu_k(\mathbf{s}_i) + \sum_{j=1}^m \lambda_{kj} f_{y,j}(\mathbf{s}_i), \quad k = 1, \dots, n_y \quad (4)$$

1
2
3
4 543 in which, compared with equation (2), the random effect $\phi_k(\mathbf{s}_i)$ is now rewritten as a truncated expansion in which
5 544 $\{f_{y,j}(\mathbf{s}_i)|j = 1, \dots, m\}$ are m common spatial factor processes underlying the n_y observed diseases and λ_{kj} is the factor
6 545 loading for variable (disease) k on the j -th factor. The factor $\mathbf{f}_y(\mathbf{s}_i) = (f_{y,1}(\mathbf{s}_i), \dots, f_{y,m}(\mathbf{s}_i))'$ is assumed to be an
7 546 m -dimensional stationary process, with $m \ll n_y$, such that $E(\mathbf{f}_y(\mathbf{s})) = \boldsymbol{\mu}_f$ and $Cov(\mathbf{f}_y(\mathbf{s})) = \boldsymbol{\Sigma}_f$.

8 In matrix form, the factor model in equation (4) can be written as

$$\boldsymbol{\eta}(\mathbf{s}_i) = \boldsymbol{\mu}(\mathbf{s}_i) + \boldsymbol{\Lambda}\mathbf{f}_y(\mathbf{s}_i),$$

9
10
11 547 where $\boldsymbol{\eta}(\mathbf{s}_i) = (\eta_1(\mathbf{s}_i), \dots, \eta_{n_y}(\mathbf{s}_i))'$ and $\boldsymbol{\Lambda} = [\lambda_{kj}]$ is a $n_y \times m$ factor-loading matrix.

12 548 Despite its simplicity, this model poses several important statistical issues. For example, though cross-loadings
13 549 are in general allowed to be estimated in $\boldsymbol{\Lambda}$, it is common practice to assume that the m underlying factors are related
14 550 only to their own manifest variables. Models with this simple structure are transparent and easily interpreted and
15 551 have also been suggested by Liu et al. (2005). Another possibility is to induce sparsity in the loading matrix. This is
16 552 especially useful in cases where n_y is large. Moreover, this choice naturally helps to overcome the drawbacks of FA,
17 553 such as unidentifiability with respect to the rotation of the latent matrices, and the difficulty of selecting the appropriate
18 554 number of factors. In fact, by imposing substantial regularization on $\boldsymbol{\Lambda}$, the identifiability issue can be alleviated when
19 555 the latent space is sufficiently sparse, and model selection criteria appears to be more effective at choosing the number
20 556 of factors because the model does not overfit to the same extent as a non-sparse model. There are currently a number
21 557 of options for how to induce sparsity constraints on the latent parameter space.

22 558 So called zero-norm priors assign finite probability mass to sparse solutions and MCMC techniques are typically
23 559 used to solve the resulting intractable inference problem (Mitchell and Beauchamp, 1988; West, 2003; Carvalho et al.,
24 560 2008). An alternative approach is to use the so called shrinkage priors which are continuous heavy-tailed densities
25 561 which favour sparse solutions. The use of shrinkage priors is more closely related to non-Bayesian sparse estimation
26 562 techniques. The canonical example is the Laplace distribution which leads to L_1 or LASSO regularisation under
27 563 Maximum a Posteriori (MAP) parameter estimation (Tibshirani, 1996; Williams, 1995). LASSO regularisation has
28 564 been used for the closely related problem of sparse principal component analysis (see, for example, Zou et al., 2006).

29 565 Shrinkage priors offer considerable computational advantages over zero-norm priors because they transform an
30 566 inference problem over discrete parameters into a continuous problem which is more easily addressed using standard
31 567 deterministic approximate inference methods (Seeger, 2008). However, although Maximum a Posteriori parameter
32 568 estimates obtained with shrinkage priors are sparse, samples from the posterior distribution will not be truly sparse.
33 569 This is a significant drawback if one is interested in characterising the uncertainty about whether or not a parameter is
34 570 exactly zero.

35 571 In the framework of hierarchical Bayesian models, it would be preferable to focus on zero-norm priors which
36 572 do assign finite probability mass to sparse solutions. These priors better characterise a prior belief in sparsity and
37 573 should therefore lead to more meaningful posterior beliefs. A natural implementation of a zero-norm sparsity prior
38 574 in this context is a spike and slab prior. This is a mixture prior on the entries of the factor loading matrix, where
39 575 one mixture component drives the weight to zero while the other mixture component allows for non-zero entries.
40 576 This prior, suggested by West (2003), not only assigns finite probability mass to truly sparse solutions, but also
41 577 allows available information about the sparse structure to be included in a natural and interpretable manner: prior
42 578 probabilities over specific entries in the mixing matrix can be used to adjust the relative weights of the corresponding
43 579 mixture components.

44 580 A further issue is that $\mathbf{f}_y(\mathbf{s})$ also requires the specification of flexible multivariate covariance structures since we
45 581 should model not only between-variable covariance but also across-location covariance. To this end, assuming a
46 582 Gaussian prior and following Liu et al. (2005), it can be shown that a rich and flexible class of variance-covariance
47 583 structures can be specified by using the linear model of coregionalization as discussed, for example, in Schmidt and
48 584 Gelfand (2003) and Banerjee et al. (2004).

49 585 Finally, we note that performing a dimensionality reduction procedure on exposure variables, as specified in
50 586 equation (4), and then regressing $\mathbf{f}_y(\mathbf{s})$ onto $\mathbf{f}_x(\mathbf{s})$ in a second level of the hierarchy, offers a possibility to overcome the
51 587 problem of multicollinearity for a set of exposure variables. As shown in Liu et al. (2005), Fontanella et al. (2015),
52 588 Ippoliti et al. (2012) and Valentini et al. (2013), this approach allows for a fully Bayesian specification of a Generalised
53 589 Structural Equation model (GSEM). In this context, a confirmatory approach, in which each variable loads only onto
54 590 a specific common factor, may facilitate the interpretation of latent variables as well as the effects of the exposure
55 591 variables on the health variables.

1
2
3
4
592 3.3. Spatio-temporal regression

5 As a result of the increasing availability of both air pollution and health data, ecological spatio-temporal studies
6 have also been developed in literature. The hierarchical models relating air pollution and health discussed above, can
7 be naturally extended to a space-time setting by using a linear function of air pollutants, covariates and space-time
8 random effects. With the conventional assumption that $Y(\mathbf{s}_i, t)$ has a Poisson distribution, *i.e.*

$$Y(\mathbf{s}_i, t) | \eta(\mathbf{s}_i, t) \stackrel{ind}{\sim} Po(E(\mathbf{s}_i, t) e^{\eta(\mathbf{s}_i, t)}),$$

9
10
11
12 593 a quite general model specifies the linear predictor as

$$\eta(\mathbf{s}_i, t) = \mu(\mathbf{s}_i, t) + \phi(\mathbf{s}_i, t) \quad (5)$$

13
14
15 where $\mu(\mathbf{s}_i, t)$ represents the large-scale spatio-temporal variability of the log-relative risk, which of course can be
16 modelled as a function of exposure variables at time t and area s_i as $\mu(\mathbf{s}_i, t) = \tilde{\mathbf{x}}(\mathbf{s}_i, t)' \boldsymbol{\beta}$. Random effects $\phi(\mathbf{s}_i, t)$ are
17 introduced to capture any residual spatio-temporal autocorrelation. A commonly used approach has been proposed by
18 Knorr-Held (2000) where a set of independent random components explains the overall risk effect over the space-time
19 domain using

$$\phi(\mathbf{s}_i, t) = \psi(\mathbf{s}_i) + \nu(\mathbf{s}_i) + \delta(t) + \zeta(t) + \theta(\mathbf{s}_i, t),$$

20
21
22
23 594 where $\nu(\mathbf{s}_i)$ and $\zeta(t)$ are independent zero mean Gaussian random effects with common variance (*i.e.* the heterogeneity
24 595 components) and $\psi(\mathbf{s}_i)$ and $\delta(t)$ are spatial and temporal random effects represented by the intrinsic CAR and first order
25 596 random walk priors, respectively. The space-time random effect $\theta(\mathbf{s}_i, t)$ has a Gaussian prior with general covariance
26 597 matrix representing cases of independence, spatial, temporal and/or spatio-temporal autocorrelation (Knorr-Held,
27 598 2000).

28 599 Another popular approach is to allow the spatially smoothed log-risk surface to evolve over time via an autore-
29 600 gressive process. Denoting with $\boldsymbol{\phi}_t = (\phi(\mathbf{s}_1, t), \dots, \phi(\mathbf{s}_{N_s}, t))'$, it may be assumed, for example, that $p(\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_T) =$
30 601 $p(\boldsymbol{\phi}_1) \prod_{t=2}^T p(\boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1})$ for which, under Gaussian priors, the marginal distribution of $\boldsymbol{\phi}_1$ has mean zero and covariance
31 602 matrix \mathbf{P}^{-1} while the conditional distributions, $\boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1}$, have conditional means which evolve in time according to a
32 603 first order autoregressive process. The matrix \mathbf{P} is a precision matrix usually specified through a CAR. This specifi-
33 604 cation is useful to model the long-term effects of air pollution on health and has been used by Ugarte et al. (2012) and
34 605 Rushworth et al. (2014). Note that an extension of this model specification to higher order autoregressive processes
35 606 or, alternatively, the use of a Space Time Autoregressive Generalised (STARG) prior (Di Giacinto et al., 2005) is in
36 607 general possible. However, their use is subject to the availability of spatial series which are rich in time.

37
38 608 Other papers dealing with spatio-temporal data include Elliott et al. (2007), Janes et al. (2007), Choi et al. (2009),
39 609 Lawson et al. (2010), Choi et al. (2011), Greven et al. (2011) and Lawson et al. (2012), with the latter being the only
40 610 paper considering confounding bias problems in space-time models.

41 611 Extending the space-time models above to a multivariate setting is not straightforward. On the other hand,
42 612 Bayesian factor models are common in multivariate time series analysis (Aguilar and West, 2000; Lopes and West,
43 613 2004) and, in this framework, the multivariate spatial structure of the geographical units suggests an extension of
44 614 the usual dynamic factor models to make use of the similarity between adjacent regions. However, the literature of
45 615 spatial dynamic factor models in health care research, especially for short-term effect studies, is still sparse and most
46 616 of the applications discuss the use of FA to achieve a dimensionality reduction of a set of covariates. An example is
47 617 offered by Reich et al. (2008) who extend the usual dynamic factor model by borrowing strength across neighboring
48 618 diameters of particles and then using the latent factors as predictors of mortality.

49 619 A new class of spatial dynamic factor models for measurements belonging to the exponential family of distribu-
50 620 tions is also discussed by Lopes et al. (2011). Though the model was developed only for one dependent variable, it
51 621 represents a direct extension to a space-time setting of the spatial factor model of equation (4). In their formulation,
52 622 the factor loadings matrix is responsible for modelling spatial variation, while the common factors capture the tempo-
53 623 ral variation. In the development of a Generalized Factor model, we use the link function to relate the conditional
54 624 mean to the linear predictor. However, in general, this does not require $\eta_k(\cdot)$ to be linear in $f_{y,j}(\cdot)$. In particular, one
55 625 could think of $\eta_k(\cdot)$ as an additive function of $f_{y,j}(\cdot)$ in the form of a generalized additive model (GAM, Hastie and
56 626 Tibshirani (1990)).

1
2
3
4 627 Assuming normality, a more structured version of the model proposed by Lopes et al. (2011) can be found in
5 628 Ippoliti et al. (2012) and Valentini et al. (2013), with the latter discussing a multivariate extension – *i.e.* $n_y > 1$.
6 629 Although this extension was not proposed in the framework of health care research, it can be a valuable tool in this
7 630 framework, offering simple solutions to the spatial misalignment problem.

9 631 **4. Available software**

11 632 Bayesian inference for the spatial and spatio-temporal models described in Section 3 is carried out using MCMC
12 633 methods (Robert and Casella, 2004; Brooks et al., 2011). A way for obtaining MCMC inference consists in running
13 634 MCMC simulations by means of hand writing code, a solution which is flexible and tailored to the specific application
14 635 but is obviously the most error-prone. Alternatively, it is possible to use a software environment designed to allow
15 636 users to perform Bayesian inference via MCMC. For example, the BUGS (Bayesian inference Using Gibbs Sampling)
16 637 project (<http://www.mrc-bsu.cam.ac.uk/software/bugs/>) has given rise to the WinBUGS (Lunn et al., 2000,
17 638 2012) software, which has opened the doors of Bayesian modelling to the wide research community. Moreover,
18 639 starting from 2004 there exists also an open-source version of the core BUGS code named OpenBUGS ([http://www.
19 640 openbugs.net/](http://www.openbugs.net/)); see Lunn et al. (2009) for a history of the BUGS project. Note that WinBUGS and OpenBUGS can
20 641 also be run within other programs like R (through the R2OpenBUGS, R2WinBUGS and BRugs packages), Stata and
21 642 SAS. Another possibility is the R CARBayes package developed by Duncan Lee (Lee, 2013) for the implementation of
22 643 Bayesian hierarchical spatial areal unit models, characterized by random effects with conditional autoregressive prior
23 644 distribution. In general the R community is extremely active in the development of new packages, listed in the CRAN
24 645 task view for Bayesian inference (<https://cran.r-project.org/web/views/Bayesian.html>) and analysis of
25 646 spatial data (<https://cran.r-project.org/web/views/Spatial.html>).

27 647 Recently, an alternative to MCMC has been proposed by Rue et al. (2009) based on the Integrated Nested Laplace
28 648 Approximation (INLA) approach, which is designed for latent Gaussian models, a wide and flexible class of models
29 649 ranging from (generalized) linear mixed to spatial and spatio-temporal models. INLA, which can be run through
30 650 the R-INLA package (<http://www.r-inla.org/>), is a deterministic algorithm for Bayesian inference and provides
31 651 accurate results in short computing time. For this reason it can be preferred when model complexity and database
32 652 dimension are troublesome from a computational point of view. Its use is now well established in several research
33 653 fields, including ecology, epidemiology, econometrics and environmental science (Rue et al., 2016); for more details
34 654 about INLA for spatial and spatio-temporal models we refer the reader to Blangiardo et al. (2013) and Blangiardo and
35 655 Cameletti (2015), where practical examples and code guidelines are also provided. Moreover, see the recent papers
36 656 Gerber and Furrer (2015) and Carroll et al. (2015) for an up-to-date comparison between INLA and MCMC methods.

39 657 **5. Discussion**

41 658 This paper has reviewed the statistical challenges involved in estimating the health impact of air pollution using
42 659 spatial and spatio-temporal ecological regression, typically based on Poisson log-linear models. An important branch
43 660 of the literature which has not been covered in this review concerns point process models to analyse case event data,
44 661 *i.e.* geo-referenced case occurrences (Diggle, 2003; Lawson, 2012). On the one hand, case event data are appealing
45 662 since they preserve exact spatial information concerning the disease occurrences, even though sometimes the use of
46 663 residential address is poorly related to individual exposure. A relevant example of point process methods in spatial
47 664 epidemiology is when case-control matched data are used to investigate the health status of people living around
48 665 potential environmental pollution sources, such as an incinerator (Diggle, 1990; Diggle et al., 2000). Controls are
49 666 matched to cases to account for heterogeneity of the underlying population at risk. In this framework, point process
50 667 methods are applied to estimate and compare the intensity surface of the point processes generating both controls and
51 668 cases, in order to understand changes in relation to the location of the putative source of hazard. In general, different
52 669 classes of point process models have been proposed in order to take into account general features of the data generating
53 670 process: the most general class for taking account of smooth spatially varying intensity is that of Log-Gaussian Cox
54 671 Processes (see Diggle (2014) for a comprehensive discussion) which allow inclusion of spatial random effects and
55 672 covariate effects. This class of models have been recently applied in epidemiological studies by Pinto Junior et al.
56 673 (2015) and Liang et al. (2008).

1
2
3
4 674 Despite the problems associated with ecological studies (as discussed in Section 2), they are far more diffused
5 675 in spatial epidemiology than studies based on case event data, not only because of their ready availability. In fact,
6 676 count data are more closely matched with the background population at risk and exposure measurements are more
7 677 reliable when they are expressed as average exposure at small area level than when they are expressed as individual
8 678 level exposure (Lawson, 2006). Generally, the more refined the spatial scale of exposure, the higher the uncertainty
9 679 associated with it; this can lead to biased estimates of the exposure-risk relationship. It should be noticed that almost
10 680 ever the researcher cannot control data aggregation since count data are collected in administrative regions, as well
11 681 as socio-demographic explanatory variables; hence ecological bias should be accounted for by means of statistical
12 682 models. One possible approach to face problems arising in ecological regression studies consists in trying to estimate
13 683 the joint distribution of response and explanatory variables within areas using a sample drawn from each area, using
14 684 the collected information to adjust the ecological regression coefficients (Plummer and Clayton, 1996). However, this
15 685 procedure is often not feasible in practice because of the high costs related to the sampling operations.

16 17 18 686 **Acknowledgements**

19 687 All the authors have been supported by the FIRB Project StEPhI (project no. RBFR12URQJ, [http://stephiproject.](http://stephiproject.it/)
20 688 [it/](http://stephiproject.it/)) provided by the Italian Ministry for Education, University and Research.

21 22 23 689 **Bibliography**

24 25 690 **References**

- 26
27 691 Agarwal, D. K., Gelfand, A. E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. *Environmental and Ecological*
28 692 *Statistics* 9 (4), 341–355.
- 29 693 Aguilar, O., West, M., 2000. Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics* 18, 338–357.
- 30 694 Atkinson, R. W., Fuller, G. W., Anderson, H. R., Harrison, R. M., Armstrong, B., 2010. Urban ambient particle metrics and health: A time-series
31 695 analysis. *Epidemiology* 21 (4).
- 32 696 Atkinson, R. W., Mills, I. C., Walton, H. A., Anderson, H. R., 2014. Fine particle components and health - systematic review and meta-analysis of
33 697 epidemiological time series studies of daily mortality and hospital admissions. *Journal of Exposure Science and Environmental Epidemiology*
34 698 25 (2), 208–214.
- 35 699 Banerjee, S., Carlin, B., Gelfand, A., 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied Probability.
700 Chapman and Hall, New York.
- 701 Basagaña, X., Aguilera, I., Rivera, M., Agis, D., Foraster, M., Marrugat, J., Elosua, R., Künzli, N., 2013. Measurement error in epidemiologic
702 studies of air pollution based on land-use regression models. *American Journal of Epidemiology* 178 (8), 1342–1346.
- 703 Bayarri, M., Berger, J., Steinberg, D., 2009. Special issue on computer modeling. *Technometrics* 51 (4), 353–353.
- 704 Bell, M. L., 2006. The use of ambient air quality modeling to estimate individual and population exposure for human health research: A case study
705 of ozone in the Northern Georgia Region of the United States. *Environment International* 32 (5), 586 – 593.
- 706 Berhane, K., Gauderman, W., Stram, D., Thomas, D., 2004. Statistical issues in studies of the long-term effects of air pollution: The Southern
707 California Children’s Health Study. *Statistical Science* 19 (3), 414–449.
- 708 Berrocal, V., Gelfand, A., Holland, D., 2011. Space-time data fusion under error in computer model output: An application to modeling air quality.
709 *Biometrics*.
- 710 Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical*
711 *Mathematics* 43 (1), 1–20.
- 712 Best, N., Richardson, S., Thomson, A., 2005. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical*
713 *Research* 14 (1), 35–59.
- 714 Biggeri, A., Bellini, P., Terracini, B., 2004. Metanalisi italiana degli studi sugli effetti a breve termine dell’inquinamento atmosferico 1996-2002.
715 *Epidemiologia e Prevenzione (Special Issue)* 28 (4-5), 1–100.
- 716 Blangiardo, M., Cameletti, M., 2015. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley.
- 717 Blangiardo, M., Cameletti, M., Baio, G., Rue, H., 2013. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiol-*
718 *ogy* 4, 33 – 49.
- 719 Blangiardo, M., Finazzi, F., Cameletti, M., 2016. Two-stage bayesian model to evaluate the effect of air pollution on chronic respiratory diseases
720 using drug prescriptions. *Spatial and Spatio-temporal Epidemiology*, –Accepted - In press.
721 URL {<http://dx.doi.org/10.1016/j.sste.2016.03.001>}
- 722 Blangiardo, M., Hansell, A., Richardson, S., 2011. A bayesian model of time activity data to investigate health effect of air pollution in time series
723 studies. *Atmospheric Environment* 45, 379–386.
- 724 Botella-Rocamora, P., Martinez-Beneito, M., Banerjee, S., 2015. A unifying modeling framework for highly multivariate disease mapping. *Statistics*
725 *in Medicine* 34 (9), 1548–1559.
- 726 Breslow, N., Day, N., 1980. *Statistical methods in cancer research*. Lyon, France, IARC Scientific Publications.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

727 Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A.,
728 Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., Kaufman, J. D., 2010. Particulate matter air pollution and cardiovascular disease: an update
729 to the scientific statement from the American Heart Association. *Circulation* 121, 12331–2378.

730 Brooks, S., Gelman, A., Jones, G., Meng, X. (Eds.), 2011. *Handbook of Markov Chain Monte Carlo*. CRC Press, Taylor & Francis Group.

731 Brown, P., 1993. *Measurement, Regression and Calibration*. Oxford University Press.

732 Brunekreef, B., Holgate, S. T., 2002. Air pollution and health. *The Lancet* 360 (9341), 1233–1242.

733 Bruno, F., Cocchi, D., Greco, F., Scardovi, E., 2014. Spatial reconstruction of rainfall fields from rain gauge and radar data. *Stochastic Environ-*
734 *mental Research and Risk Assessment* 28 (5), 1235–1245.

735 Bruno, F., Paci, L., 2013. Spatio-temporal model for short-term predictions of air pollution data. In: Lanzarone, E., Ieva, F. (Eds.), *The contribution*
736 *of Young Researchers to Bayesian Statistics - Proceedings of BAYSM2013*. Milan, 5-6 June, 2013.

737 Cameletti, M., Ignaccolo, R., Bande, S., 2011. Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics* 22, 985–996.

738 Carey, I. M., Atkinson, R. W., Kent, A. J., van Staa, T., Cook, D. G., Anderson, H. R., 2013. Mortality associations with long-term exposure to
739 outdoor air pollution in a national english cohort. *American Journal of Respiratory and Critical Care Medicine* 187 (11), 1226–1233.

740 Carlin, B., Banerjee, S., 2003. Hierarchical multivariate car models for spatio-temporally correlated survival data (with discussion). In: Bernardo,
741 J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M. (Eds.), *Bayesian Statistics 7*. Oxford University Press, London, pp.
742 45–63.

743 Carlin, B., Xia, H., Devine, O., Tolbert, P., Mulholland, J., 1999. Spatio-Temporal Hierarchical Models for Analyzing Atlanta Pediatric Asthma ER
744 Visit Rates. In: Gatsonis, C., Kass, R., Carlin, B., Carriquiry, A., Gelman, A., Verdinelli, I., West, M. (Eds.), *Case Studies in Bayesian Statistics*.
745 Vol. 140 of *Lecture Notes in Statistics*. Springer New York, pp. 303–320.

746 Carpenter, T., 2011. The spatial epidemiologic (r)evolution: A look back in time and forward to the future. *Spatial and Spatio-temporal Epidemi-*
747 *ology* 2 (3), 119 – 124, special Issue: {GEOVET} 2010.

748 Carroll, R., Lawson, A., Faes, C., Kirby, R., Aregay, M., Watjou, K., 2015. Comparing INLA and OpenBUGS for hierarchical Poisson modeling
749 in disease mapping. *Spatial and Spatio-temporal Epidemiology* 14-15, 45 – 54.

750 Carroll, R., Ruppert, D., Stefanski, L., Crainiceanu, C., 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edn. Chapman
751 and Hall/CRC: Boca Raton, FL.

752 Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., West, M., 2008. High-dimensional sparse factor modeling: Applications in gene expression
753 genomics. *Journal of the American Statistical Association* 103 (484), 1438–1456.

754 Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., Forastiere, F., 2013. Long-term exposure to urban air pollution and
755 mortality in a cohort of more than a million adults in rome. *Environmental Health Perspectives* 121 (3), 324–331.

756 Choi, J., Fuentes, M., Reich, B. J., 2009. Spatial-temporal association between fine particulate matter and daily mortality. *Computational Statistics*
757 *& Data Analysis* 53 (8), 2989 – 3000.

758 Choi, J., Lawson, A., Cai, B., Hossain, M., 2011. Evaluation of Bayesian spatial-temporal latent models in small area health data. *Environmetrics*
759 22 (8), 1008–1022.

760 Clayton, D., Bernardinelli, L., Montomoli, C., 1993. Spatial correlation in ecological analysis. *International Journal of Epidemiology* 22, 1193–
761 1202.

762 Cocchi, D., Greco, F., Trivisano, C., 2007. Hierarchical space-time modelling of PM10 pollution. *Atmospheric Environment* 41, 532–542.

763 Cressie, N., 1993. *Statistics for Spatial Data*. Wiley.

764 Di Giacinto, V., Dryden, I., Ippoliti, L., Romagnoli, L., 2005. Linear smoothing of noisy spatial temporal series. *Journal of Mathematics and*
765 *Statistics* 1 (4), 300–312.

766 Diggle, P., 1990. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of*
767 *the Royal Statistical Society. Series A (Statistics in Society)* 153 (3), 349–362.

768 Diggle, P., 2003. *Statistical Analysis of Spatial Point Patterns*. Mathematics in biology. Arnold.

769 Diggle, P., 2014. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, Third Edition. Chapman and Hall/CRC, New York.

770 Diggle, P. J., Morris, S. E., Wakefield, J. C., 2000. Point-source modelling using matched case-control data. *Biostatistics* 1 (1), 89–105.
771 URL <http://biostatistics.oxfordjournals.org/content/1/1/89.abstract>

772 Dobra, A., Lenkoski, A., Rodriguez, A., 2011. Bayesian inference for general gaussian graphical models with application to multivariate lattice
773 data. *Journal of the American Statistical Association* 106 (496), 1418–1433.

774 Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G., Speizer, F. E., 1993. An association between air
775 pollution and mortality in six US cities. *New England journal of medicine* 329 (24), 1753–1759.

776 Eberly, L. E., Carlin, B. P., 2000. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in*
777 *Medicine* 19 (17-18), 2279–2294.

778 Elgethun, K., Fenske, R., Yost, M., Oalcisko, G., 2003. Time-location analysis for exposure assessment studies of children using a novel global
779 positioning system instrument. *Environ Health Perspect* 111, 115–122.

780 Elliott, P., Shaddick, G., Wakefield, J. C., Hoogh, C. d., Briggs, D. J., 2007. Long-term associations of outdoor air pollution with mortality in Great
781 Britain. *Thorax* 62 (12), 1088–1094.

782 Elliott, P., Wartenberg, D., 2004. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives* 112 (9),
783 998–1006.

784 Evensen, G., 2009. *Data assimilation. The Ensemble Kalman Filter*. Springer.

785 Fassò, A., Cameletti, M., 2009. The EM algorithm in a distributed computing environment for modelling environmental space-time data. *Environ-*
786 *mental Modelling & Software* 24, 1027–1035.

787 Fassò, A., Cameletti, M., Nicolis, O., 2007. Air quality monitoring using heterogeneous networks. *Environmetrics* 18, 245–264.

788 Fassò, A., Finazzi, F., 2011. Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics* 22,
789 735–748.

790 Fontanella, L., Ippoliti, L., Sarra, A., Valentini, P., Palermi, S., 2015. Hierarchical generalised latent spatial quantile regression models with
791 applications to indoor radon concentration. *Stochastic Environmental Research and Risk Assessment* 29 (2), 357–367.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

792 Fuentes, M., Raftery, A., 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical
793 models. *Biometrics* 61, 36–45.

794 Fuentes, M., Song, H.-R., Ghosh, S. K., Holland, D. M., Davis, J. M., 2006. Spatial association between speciated fine particles and mortality.
795 *Biometrics* 62 (3), 855–863.

796 Fuller, W., 1987. *Measurement Error Models*. New York: John Wiley & Sons.

797 Gelfand, A., 2012. Hierarchical modeling for spatial data problems. *Spatial Statistics* 1, 30 – 39.

798 Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P. (Eds.), 2010. *Handbook of Spatial Statistics*. Chapman & Hall.

799 Gelfand, A., Vounatsou, P., 2003. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 4 (1), 11–15.

800 Gerber, F., Furrer, R., 2015. Pitfalls in the Implementation of Bayesian Hierarchical Modeling of Areal Count Data: An Illustration Using BYM
801 and Leroux Models. *Journal of Statistical Software* 63 (1).

802 Gerharz, L., Krüger, A., Klemm, O., 2009. Applying indoor and outdoor modeling techniques to estimate individual exposure to PM2.5 from
803 personal GPS profiles and diaries: a pilot study. *Science of the Total Environment* 407, 5184–5193.

804 Gotway, C., Young, L., 2002. Combining incompatible spatial data. *Journal of the American Statistical Association* 97 (458), 632–648.

805 Gray, S., Gelfand, A., Miranda, M., 2011. Hierarchical spatial modeling of uncertainty in air pollution and birth weight study. *Statistics in Medicine*
806 30 (17), 2187–2198.

807 Greco, F. P., Trivisano, C., 2009. A multivariate CAR model for improving the estimation of relative risks. *Statistics in Medicine* 28 (12), 1707–
808 1724.

809 Greenland, S., Morgenstern, H., 1989. Ecological bias, confounding and effects modification. *International Journal of Epidemiology* 18, 269–274.

810 Greenland, S., Robins, J., 1994. Invited commentary: ecologic studies - biases, misconceptions, and counterexamples. *American Journal of Epi-
811 demiology* 139, 747–760.

812 Grell, K., Diggle, P., Frederiksen, K., Schüz, J., Cardis, E., Andersen, P., 2015. A three-dimensional point process model for the spatial distribution
813 of disease occurrence in relation to an exposure source. *Statistics in Medicine* 34, 3170–3180.

814 Greven, S., Dominici, F., Zeger, S., 2011. An approach to the estimation of chronic air pollution effects using spatio-temporal information. *Journal
815 of the American Statistical Association* 106 (494), 396–406.

816 Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., Coull, B. A., 2009. Measurement error caused by spatial misalignment in environmental
817 epidemiology. *Biostatistics* 10 (2), 258–274.

818 Gschlößl, S., Czado, C., 2008. Modelling count data with overdispersion and spatial effects. *Statistical Papers* 49 (3), 531–552.

819 Hastie, T., Tibshirani, R., 1990. *Generalized Additive Models*. Chapman & Hall/CRC.

820 Hodges, J. S., Reich, B. J., 2010. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* 64 (4),
821 325–334.

822 Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess
823 spatial variation of outdoor air pollution. *Atmospheric Environment* 42 (33), 7561 – 7578.

824 Hoek, G., Krishnan, R., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., Kaufman, J., 2013. Long-term air pollution exposure and cardio-
825 respiratory mortality: a review. *Environmental Health* 12 (1), 12–43.

826 Huang, G., Lee, D., Scott, M., 2015. An integrated Bayesian model for estimating the long-term health effects of air pollution by fusing modelled
827 and measured pollution data: A case study of nitrogen dioxide concentrations in Scotland. *Spatial and Spatio-temporal Epidemiology* 14-15, 63
828 – 74.

829 Hubbell, B. J., Hallberg, A., McCubbin, D. R., Post, E., 2005. Health-related benefits of attaining the 8-hr ozone standard. *Environmental Health
830 Perspectives* 113 (1), 73–82.

831 Hughes, J., Haran, M., 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal
832 Statistical Society: Series B (Statistical Methodology)* 75 (1), 139–159.

833 Ignaccolo, R., Ghigo, S., Bande, S., 2013. Functional zoning for air quality. *Environmental and Ecological Statistics* 20, 109–127.

834 Ignaccolo, R., Mateu, J., Giraldo, R., 2014. Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental
835 Research and Risk Assessment* 28 (5), 1171–1186.

836 Ippoliti, L., Martin, R., L., R., 2015. A note on efficient likelihood computations for some multivariate Gaussian Markov Random Fields. Preprint
837 available from the authors, *Submitted*.

838 Ippoliti, L., Valentini, P., Gamerman, D., 2012. Space-time modelling of coupled spatio-temporal environmental variables. *Journal of the Royal
839 Statistical Society, Series C* 61, 175–200.

840 Jackson, C., Best, N., Richardson, S., 2008. Hierarchical related regression for combining aggregate and individual data in studies of socio-
841 economic disease risk factors. *Statistics in Medicine* 171.

842 Janes, H., Dominici, F., Zeger, S., 2007. Trends in air pollution and mortality: an approach to the assessment of unmeasured confounding.
843 *Epidemiology* 18 (4), 416–23.

844 Jerrett, M., Gale, S., Kontgis, C., 2010. *Spatial Modeling in Environmental and Public Health Research*. *International Journal of Environmental
845 Research and Public Health* 7 (4), 1302–1329.

846 Jin, X., Banerjee, S., Carlin, B. P., 2007. Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the
847 Royal Statistical Society: Series B (Statistical Methodology)* 69 (5), 817–838.

848 Jin, X., Carlin, B. P., Banerjee, S., 2005. Generalized hierarchical multivariate CAR models for areal data. *Biometrics* 61 (4), 950–961.

849 Kelsall, J., Wakefield, J., 2002. Modeling spatial variation in disease risk. *Journal of the American Statistical Association* 97 (459), 692–701.

850 Kim, H., Sun, D., Tsutakawa, R., 2001. A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional
851 autoregressive model. *Journal of the American Statistical Association* 96, 1506–1521.

852 Knorr-Held, L., 2000. Bayesian modelling of inseparable space-time variation in disease risk. *Environmetrics* 19 (17), 2555–67.

853 Krall, J. R., Chang, H. H., Sarnat, S. E., Peng, R. D., Waller, L. A., 2015. Current methods and challenges for epidemiological studies of the
854 associations between chemical constituents of particulate matter and health. *Current Environmental Health Reports* 2 (4), 388–398.

855 Künzli, N., 2012. Is air pollution of the 20th century a cause of current asthma hospitalisations? *Thorax* 67 (1), 2–3.

856 Lanki, T., Pekkanen, J., Aalto, P., Elosua, R., Berglind, N., D’Ippoliti, D., Kulmala, M., Nyberg, F., Peters, A., Picciotto, S., Salomaa, V., Sunyer,

- 1
2
3
4 857 J., Tiittanen, P., von Klot, S., Forastiere, F., 2006. Associations of traffic related air pollutants with hospitalisation for first acute myocardial
5 858 infarction: the HEAPSS study. *Occupational and environmental medicine* 63 (12), 844–851.
- 6 859 Lawson, A., 2006. *Statistical Methods in Spatial Epidemiology*. Wiley.
- 7 860 Lawson, A., 2009. *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*. CRC Press.
- 8 861 Lawson, A., Choi, J., Cai, B., Hossain, M., Kirby, R., Liu, J., 2012. Bayesian 2-Stage Space-Time Mixture Modeling With Spatial Misalignment
9 862 of the Exposure in Small Area Health Data. *Journal of Agricultural, Biological, and Environmental Statistics* 17 (3), 417–441.
- 10 863 Lawson, A., Song, H., Cai, B., Hossain, M., Huang, K., 2010. Space-time latent component modeling of geo-referenced health data. *Statistics in
11 864 Medicine* 29 (17), 2012–2027.
- 12 865 Lawson, A. B., 2012. Bayesian point event modeling in spatial and environmental epidemiology. *Statistical Methods in Medical Research* 21 (5),
13 866 509–529.
- 14 867 URL <http://smm.sagepub.com/content/21/5/509.abstract>
- 15 868 Lee, A., Szpiro, A., Kim, S., Sheppard, L., 2015. Impact of preferential sampling on exposure prediction and health effect inference in the context
16 869 of air pollution epidemiology. *Environmetrics* 26 (4), 255–267.
- 17 870 Lee, D., 2013. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*
18 871 55 (13), 1–24.
- 19 872 Lee, D., Rushworth, A., Sahu, S. K., 2014. A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution.
20 873 *Biometrics* 70 (2), 419–429.
- 21 874 Lee, D., Sarran, C., 2015. Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Envi-
22 875 ronmetrics* 26 (7), 477–487.
- 23 876 Lee, D., Shaddick, G., 2010. Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics* 66 (4), 1238–1246.
- 24 877 Lee, S.-J., Yeatts, K. B., Serre, M. L., 2009. A Bayesian Maximum Entropy approach to address the change of support problem in the spatial
25 878 analysis of childhood asthma prevalence across North Carolina. *Spatial and Spatio-temporal Epidemiology* 1 (1), 49 – 60.
- 26 879 Lepeule, J., Laden, F., Dockery, D., Schwartz, J., 2012. Chronic exposure to fine particles and mortality: An extended follow-up of the harvard six
27 880 cities study from 1974 to 2009. *Environmental Health Perspectives* 120 (7), 965–970.
- 28 881 Li, L., Lossner, T., Yorke, C., Piltner, R., 2014. Fast inverse distance weighting-based spatiotemporal interpolation: A web-based application of
29 882 interpolating daily fine particulate matter PM2.5 in the contiguous U.S. using parallel programming and k-d tree. *International Journal of
30 883 Environmental Research and Public Health* 11 (9), 9101.
- 31 884 Liang, S., Carlin, B., Gelfand, A., 2008. Analysis of Minnesota Colon and Rectum cancer point patterns with spatial and nonspatial covariate
32 885 information. *The annals of applied statistics* 3 (3), 943–962.
- 33 886 Liu, X., Wall, M. M., Hodges, J. S., 2005. Generalized spatial structural equation models. *Biostatistics* 6 (4), 539–557.
- 34 887 Lopes, H., Gamerman, D., Salazar, E., 2011. Generalized spatial dynamic factor models. *Computational Statistics and Data Analysis* 55, 1319–
35 888 1330.
- 36 889 Lopes, H., West, M., 2004. Bayesian model assessment in factor analysis. *Statistica Sinica* 14, 41–67.
- 37 890 Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D., 2012. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Press.
- 38 891 Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*
39 892 25 (28), 3049–3067.
- 40 893 Lunn, D., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility.
41 894 *Statistics and Computing* 10 (4), 325–337.
- 42 895 MacNab, Y. C., 2003. Hierarchical bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics* 59 (2),
43 896 305–315.
- 44 897 MacNab, Y. C., 2011. On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research* 20 (1), 49–68.
- 45 898 Madsen, L., Ruppert, D., Altman, N. S., 2008. Regression with spatially misaligned data. *Environmetrics* 19 (5), 453–467.
- 46 899 Mardia, K., 1988. Multi-dimensional multivariate gaussian markov random fields with applications to image processing. *Journal of Multivariate
47 900 Analysis* 24, 265–284.
- 48 901 Martinez-Beneito, M., 2013. A general modelling framework for multivariate disease mapping. *Biometrika* 100, 539–553.
- 49 902 McBride, S. J., Williams, R. W., Creason, J., 2007. Bayesian hierarchical modeling of personal exposure to particulate matter. *Atmospheric
50 903 Environment* 41 (29), 6143 – 6155.
- 51 904 Meliker, J., Sloan, C., 2011. Spatio-temporal epidemiology: Principles and opportunities. *Spatial and Spatio-temporal Epidemiology* 2, 1–9.
- 52 905 Mitchell, T., Beauchamp, J., 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83, 1023–1032.
- 53 906 Molitor, J., Molitor, N.-T., Jerrett, M., McConnell, R., Gauderman, J., Berhane, K., Thomas, D., 2006. Bayesian modeling of air pollution health
54 907 effects with missing exposure data. *American Journal of Epidemiology* 164 (1), 69–76.
- 55 908 Morgenstern, H., 1998. *Encyclopedia of Biostatistics*. Vol. 2. JohnWiley & Sons, New York, Ch. Ecologic study, pp. 1255–1276.
- 56 909 Nuckols, J. R., Ward, M. H., Jarup, L., 06 2004. Using geographic information systems for exposure assessment in environmental epidemiology
57 910 studies. *Environmental Health Perspectives* 112 (9), 1007–1015.
- 58 911 Openshaw, S., Taylor, P., 1981. *Quantitative geography*. London. Routledge, Ch. The modifiable areal unit problem.
- 59 912 Osei, F., 2014. Current statistical methods for spatial epidemiology: A review. *Austin Biometrics and Biostatistics* 1 (2), 1–7.
- 60 913 Ozkaynak, H., Baxter, L. K., Dionisio, K. L., Burke, J., 2013. Air pollution exposure prediction approaches used in air pollution epidemiology
61 914 studies. *Journal of Exposure Science and Environmental Epidemiology* 23 (6), 566–572.
- 62 915 Paci, L., Gelfand, A., Cocchi, D., 2015. Quantifying uncertainty for temperature maps derived from computer models. *Spatial Statistics* 12, 96–108.
- 63 916 Paci, L., Gelfand, A., Holland, M., 2013. Spatio-temporal modeling for real-time ozone forecasting. *Spatial Statistics* 4, 79–93.
- 64 917 Paciorek, C. J., 2010. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical science: a
65 918 review journal of the Institute of Mathematical Statistics* 25 (1), 107–125.
- 919 Peng, R., Dominici, F., 2008. *Statistical Methods for Environmental Epidemiology with R. A Case Study in Air Pollution and Health*. Springer.
- 920 Peng, R. D., Bell, M. L., 2010. Spatial misalignment in time series studies of air pollution and health data. *Biostatistics (Oxford, England)* 11 (4),
921 720–740.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

922 Peng, R. D., Dominici, F., Louis, T. A., 2006. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169 (2), 179–203.

923

924 Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L., Samet, J. M., 2005. Seasonal Analyses of Air Pollution and Mortality in 100 US Cities. *American Journal of Epidemiology* 161 (6), 585–594.

925

926 Peng, R. D., Dominici, F., Welty, L. J., 2009. A bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58 (1), 3–24.

927

928 Piantadosi, S., Byar, D., Green, S., 1988. The ecological fallacy. *American Journal of Epidemiology* 127 (5), 893–904.

929

930 Pinto Junior, J. A., Gamerman, D., Paez, M. S., Fonseca Alves, R. H., 2015. Point pattern analysis with spatially varying covariate effects, applied to the study of cerebrovascular deaths. *Statistics in Medicine* 34 (7), 1214–1226.

931

932 Pirani, M., Best, N., Blangiardo, M., Liverani, S., Atkinson, R. W., Fuller, G. W., 2015. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International* 79, 56 – 64.

933

934 Plummer, M., Clayton, D., 1996. Estimation of population exposure in ecological studies. *Journal of the Royal Statistical Society, Series B* 58, 113–126.

935

936 Reich, B., Fuentes, M., Burke, J., 2008. Analysis of the effects of ultrafine particulate matter while accounting for human exposure. *Environmetrics* 20 (2), 131–146.

937

938 Reich, B. J., Hodges, J. S., Zadnik, V., 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* 62 (4), 1197–1206.

939

940 Roberston, W., 1950. Ecological correlation and the behaviour of individuals. *American Sociological Review* 15, 351–357.

941

942 Robert, C., Casella, G., 2004. *Monte Carlo Statistical Methods*. Springer.

943

944 Robertson, C., Nelson, T. A., MacNab, Y. C., Lawson, A. B., 2010. Review of methods for space-time disease surveillance. *Spatial and Spatio-temporal Epidemiology* 1 (2-3), 105 – 116, GEOMED Conference.

945

946 Rothman, K., Lash, T., Greenland, S., 2012. *Modern Epidemiology*, 3rd Edition. Wolters Kluwer.

947

948 Rue, H., Held, L., 2005. *Gaussian Markov Random Fields. Theory and Applications*. Chapman & Hall.

949

950 Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2), 319–392.

951

952 Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., Lindgren, F. K., 2016. Bayesian Computing with INLA: A Review. ArXiv e-prints. URL <http://adsabs.harvard.edu/abs/2016arXiv160400860R>

953

954 Rushworth, A., Lee, D., Mitchell, R., 2014. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology* 10, 29–38.

955

956 Sahu, S., Gelfand, A., Holland, D., 2006. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological and Environmental Statistics* 11, 61–86.

957

958 Sahu, S., Mardia, K., 2005. A Bayesian Kriged-Kalman model for short-term forecasting of air pollution level. *Journal of the Royal Statistical Society, Series C* 54, 223–244.

959

960 Sain, S., Cressie, N., 2007. A spatial model for multivariate lattice data. *Journal of Econometrics* 140, 226–259.

961

962 Sain, S., Furrer, R., Cressie, N., 2011. A spatial analysis of multivariate output from regional climate models. *The Annals of Applied Statistics* 5, 150–175.

963

964 Samet, M., Zeger, S., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J., Zanobetti, A., 2000. *The National Morbidity, Mortality, and Air Pollution Study*. Tech. rep., Research Report: Health Effect Institute, Cambridge, MA.

965

966 Samoli, E., Touloumi, G., Zanobetti, A., Le Tertre, A., Schindler, C., Atkinson, R., Vonk, J., Rossi, G., Saez, M., Rabczenko, D., Schwartz, J., Katsouyanni, K., 2003. Investigating the dose-response relation between air pollution and total mortality in the aphea-2 multicity project. *Occupational and Environmental Medicine* 60 (12), 977–982.

967

968 Schikowski, T., Sugiri, D., Ranft, U., Gehring, U., Heinrich, J., Wichmann, H.-E., Kramer, U., 2005. Long-term air pollution exposure and living close to busy roads are associated with COPD in women. *Respiratory Research* 6 (1), 152.

969

970 Schmidt, A., Gelfand, A., 2003. A bayesian coregionalization model for multivariate pollutant data. *Journal of Geophysical Research* 108 (494), 8783.

971

972 Schwartz, J., Marcus, A., 1990. Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology* 131 (1), 185–194.

973

974 Seeger, M., 2008. Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research* 9, 759–813.

975

976 Shaddick, G., Zidek, J., 2014. A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spatial Statistics* 9, 51–65.

977

978 Shaddick, G., Zidek, J., 2015. *Spatio-Temporal Methods in Environmental Epidemiology*. Chapman & Hall.

979

980 Sheppard, L., Burnett, R., Szpiro, A., Kim, S., Jerrett, M., Pope, C., Brunekreef, B., 2012. Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere & Health* 5 (2), 203–216.

981

982 Smith, R., Davis, J., Sacks, J., Speckman, P., Styer, P., 2000. Regression models for air pollution and daily mortality: Analysis of data from Birmingham, Alabama. *Environmetrics* 11, 719–734.

983

984 Smith, R., Kolenikov, S., 2003. Spatiotemporal modeling of PM_{2.5} data with missing values. *Journal of Geophysical Research* 108 (D24), 11–1,11–11.

985

986 Stern, H., Cressie, N., 2000. Posterior predictive model checks for disease mapping models. *Statistics in Medicine* 19, 2377–97.

987

988 Szpiro, A., Sheppard, L., Lumley, T., 2011. Efficient measurement error correction with spatially misaligned data. *Biostatistics* 12 (4), 610–623.

989

990 Szpiro, A. A., Paciorek, C. J., 2013. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* 24 (8), 501–517.

991

992 Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58, 267–288.

993

994 Ugarte, M., Etxebarria, J., Goicoa, T., Ardanaz, E., 2012. Gender-specific spatiotemporal patterns of colorectal cancer incidence in Navarre, Spain (1990-2005). *Cancer Epidemiology* 36 (3), 254–62.

995

996 Valentini, P., Ippoliti, L., Fontanella, L., 2013. Modeling us housing prices by spatial dynamic structural equation models. *The Annals of Applied Statistics* 7, 763–798.

997

998 Van de Kasstele, J., Stein, A., 2006. A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

987 model output. *Environmetrics* 17, 309–322.

988 Van de Kasstele, J., Stein, A., Dekkers, A., Velders, G., 2009. External drift kriging of nox concentrations with dispersion model output in a
989 reduced air quality monitoring network. *Environ Ecol Stat* 16 (3), 321–339.

990 Wakefield, J., 2003. A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics* 11,
991 31–54.

992 Wakefield, J., 2007. Disease mapping and spatial regression with count data. *Biostatistics* 8 (2), 158–183.

993 Wakefield, J., Elliot, P., 1999. Issues in statistical analysis of small area health data. *Statistics in Medicine* 18, 2377–2399.

994 Wakefield, J., Lyons, H., 2010. *Handbook of Spatial Statistics*. Chapman & Hall, Ch. Spatial aggregation and the ecological fallacy.

995 Wakefield, J., Shaddick, G., 2006. Health-exposure modeling and the ecological fallacy. *Biostatistics* 7 (3), 438–455.

996 Waller, L., Gotway, C., 2004. *Applied Spatial Statistics for Public Health Data*. Wiley.

997 West, M., 2003. Bayesian factor regression models in the “Large p, Small n” paradigm. *Bayesian Statistics* 7, 723–732.

998 WHO, 2015. Economic cost of the health impact of air pollution in Europe: Clean air, health and wealth. Tech. rep., WHO Regional Of-
999 fice for Europe, [http://www.euro.who.int/en/media-centre/events/events/2015/04/ehp-mid-term-review/publications/](http://www.euro.who.int/en/media-centre/events/events/2015/04/ehp-mid-term-review/publications/economic-cost-of-the-health-impact-of-air-pollution-in-europe)
1000 [economic-cost-of-the-health-impact-of-air-pollution-in-europe](http://www.euro.who.int/en/media-centre/events/events/2015/04/ehp-mid-term-review/publications/economic-cost-of-the-health-impact-of-air-pollution-in-europe).

1001 Wikle, C., Berliner, L., 2005. Combining information across spatial scales. *Technometrics* 47 (1), 80–91.

1002 Williams, P., 1995. Bayesian regularization and pruning using a Laplace prior. *Neural Computation* 7, 117–143.

1003 Young, L. J., Gotway, C. A., Yang, J., Kearney, G., DuClos, C., 2009. Linking health and environmental data in geographical analysis: It’s so much
1004 more than centroids. *Spatial and Spatio-temporal Epidemiology* 1 (1), 73 – 84.

1005 Zhang, Y., Hodges, J., Banerjee, S., 2009. Smoothed ANOVA with spatial effects as a competitor to MCAR in multivariate spatial smoothing.
1006 *Annals of Applied Statistics* 3 (4), 1805–1830.

1007 Zhu, L., Carlin, B. P., Gelfand, A., 2003. Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER
1008 visits in Atlanta. *Environmetrics* 14 (5), 537–557.

1009 Zidek, J., Shaddick, G., Taylor, C., 2014. Reducing estimation bias in adaptively changing monitoring networks with preferential site selection.
1010 *The Annals of Applied Statistics* 8 (3), 1640–1670.

1011 Zidek, J., Sun, L., Le, N., Ozkaynak, H., 2002. Contending with space-time interaction in the spatial prediction of pollution: Vancouver’s hourly
1012 ambient PM₁₀ field. *Environmetrics* 13, 595–613.

1013 Zidek, J. V., Shaddick, G., Meloche, J., Chatfield, C., White, R., 2007. A framework for predicting personal exposures to environmental hazards.
1014 *Environmental and Ecological Statistics* 14 (4), 411–431.

1015 Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 1, 265–286.