Water and Soil Pollution: Ecological Environmental Study Methodologies Useful for Public Health Projects. A Literature Review

Roberto Lillini, Andrea Tittarelli, Martina Bertoldi, David Ritchie, Alexander Katalinic, Ron Pritzkuleit, Guy Launoy, Ludivine Launay, Elodie Guillaume, Tina Žagar, Carlo Modonesi, Elisabetta Meneghini, Camilla Amati, Francesca Di Salvo, Paolo Contiero, Alessandro Borgini, and Paolo Baili

Contents

WASABY Working Group have been listed in the acknowledgments.

Roberto Lillini, Andrea Tittarelli, Alessandro Borgini, and Paolo Baili contributed equally to this work.

R. Lillini (✉) · E. Meneghini · C. Amati · P. Baili
Analytical Epidemiology and Health Impact Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
e-mail: roberto.lillini@istitutotumori.mi.it; elisabetta.meneghini@istitutotumori.mi.it; lifetable@istitutotumori.mi.it

A. Tittarelli
Cancer Registry Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
e-mail: andrea.tittarelli@istitutotumori.mi.it

M. Bertoldi · P. Contiero
Environmental Epidemiology Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
e-mail: martina.bertoldi@istitutotumori.mi.it; paolo.contiero@istitutotumori.mi.it

D. Ritchie
Association Européenne des Ligues contre le Cancer, Bruxelles, Belgium
e-mail: david@europeancancerleagues.org

P. de Voogt (ed.), *Reviews of Environmental Contamination and Toxicology Volume 256*,
Reviews of Environmental Contamination and Toxicology 256,
https://doi.org/10.1007/398_2020_58

Abstract Health risks at population level may be investigated with different types of environmental studies depending on access to data and funds. Options include ecological studies, case–control studies with individual interviews and human sample analysis, risk assessment or cohort studies. Most public health projects use data and methodologies already available due to the cost of ad-hoc data collection. The aim of the article is to perform a literature review of environmental exposure and health outcomes with main focus on methodologies for assessing an association between water and/or soil pollutants and cancer. A systematic literature search was performed in May 2019 using PubMed. Articles were assessed by four independent reviewers. Forty articles were identified and divided into four groups, according to the data and methods they used, i.e.: (1) regression models with data by geographical area; (2) regression models with data at individual level; (3) exposure intensity threshold values for evaluating health outcome trends; (4) analyses of distance between source of pollutant and health outcome clusters. The issue of exposure

A. Katalinic · R. Pritzkuleit
Institute for Cancer Epidemiology at the University Lübeck, Lübeck, Germany
e-mail: Alexander.Katalinic@uksh.de; Ron.Pritzkuleit@uksh.de

G. Launoy
Normandie Univ, UNICAEN, INSERM, ANTICIPE, Caen, France

Pôle recherche – Centre Hospitalier Universitaire, Caen, France
e-mail: guy.launoy@unicaen.fr

L. Launay
Normandie Univ, UNICAEN, INSERM, ANTICIPE, Caen, France

Centre François Baclesse, Caen, France
e-mail: ludivine.launay@inserm.fr

E. Guillaume
Normandie Univ, UNICAEN, INSERM, ANTICIPE, Caen, France
e-mail: elodie.guillaume@unicaen.fr

T. Žagar
Institute of Oncology Ljubljana, Ljubljana, Slovenia
e-mail: TZagar@onko-i.si

C. Modonesi
Cancer Registry Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy

International Society of Doctors for the Environment (ISDE), Arezzo, Italy
e-mail: carlo.modonesi@istitutotumori.mi.it

F. Di Salvo
Pancreas Translational and Clinical Research Center, Ospedale IRCCS "San Raffaele", Milan, Italy
e-mail: disalvo.francesca@hsr.it

A. Borgini
Environmental Epidemiology Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy

International Society of Doctors for the Environment (ISDE), Arezzo, Italy
e-mail: alessandro.borgini@istitutotumori.mi.it

assessment has been investigated for over 40 years and the most important innovations regard technologies developed to measure pollutants, statistical methodologies to assess exposure, and software development. Thanks to these changes, it has been possible to develop and apply geo-coding and statistical methods to reduce the ecological bias when considering the relationship between humans, geographic areas, pollutants, and health outcomes. The results of the present review may contribute to optimize the use of public health resources.

Keywords Health outcomes · Public health · Soil pollution · Spatial analysis · Statistical methods · Water pollution

1 Introduction

The effects of the industrialization of many economic activities during the last two centuries have become an important issue for environmental and human health. As recognized by the World Health Organization (WHO), environmental integrity is a major determinant of health and actions to protect human and animal populations from diseases should be a primary objective of a global health agenda, particularly in the case of degenerative pathologies. Based on the above principles, the European health policy framework "Health 2020", supported by WHO's Regional Office for Europe, aims to improve the health and well-being of European citizens by reducing the weight of all disease determinants (World Health Organization Website 2019). Among actions promoted by Health 2020, management and remediation policies for preserving the resilient functions of ecosystems and environmental matrices are crucial.

A challenging issue lies in the considerable health risk resulting from the exposure to toxic chemicals and other stressors of industrial origin, as documented by a number of studies developed also in Europe (Hänninen et al. 2014). This kind of investigation can prove problematic, as people are exposed to hundreds of toxicants that come from both anthropogenic and natural sources: their physical and chemical interactions determine an extremely complex picture of phenomena that must necessarily to be taken into account. Contaminants move across environmental matrices and often accumulate in the organisms therein. The assessment of potential health effects due to exposure to all factors is a demanding task, often one too complex to be performed. Some chemicals are widespread on a global scale, while others accumulate around industrial or other specific sites; in this case, their concentration significantly exceeds that of background values. This results in considerable disparities in the level of exposure of human populations (Stewart and Wild 2014) and increases the obstacles when trying to explore the relationship between environmental pollution and health outcomes. However, an appropriate epidemiologic approach can contribute to clarify causes of disease, factors conferring susceptibility, and actual levels of exposure at which health effects occur (Deener et al. 2018).

Health risks at population level may be investigated with different types of environmental studies depending on access to data and funds. Options include ecological studies, case–control studies with individual interviews and human sample analysis, risk assessment or cohort studies. (Baker et al. 1999).

In 2016, the Health and Food Safety Directorate General (DG SANTE) of the European Commission launched, under the 3rd Health Programme, a call for project proposals aiming to identify geographical regions presenting higher breast cancer rates within the European Union, and to investigate the statistical correlation between water and soil polluting agents and high cancer rates (European Commission 2016). The *WASABY – Water And Soil contamination and Awareness on Breast cancer risk in Young women* project was established with the following objectives: (1) mapping breast cancer risk to identify areas at higher risk using specific geographic information systems; (2) reviewing scientific literature on the relationship between water and soil pollutants and breast cancer risk, and on possible methods for a pilot ecological environmental study (WASABY Website 2019). We defined the above objectives in consideration of the scopes of a DG SANTE call (i.e., excluding analytic studies such as cohort or case–control studies which could aim to evaluate cause-and-effect relationship) and of the call budget (European Commission 2016).

As most public health projects, WASABY focuses its activities using available data and methodologies (i.e., incidence data from cancer registries and databases of environmental agencies, spatial mapping methods, and ecological regression methods used in environmental studies).

In the present article, we summarize a PubMed (National Center for Biotechnology Information 2019) literature review of methodologies applied across the world to study the correlation between water and soil pollutants (e.g., arsenic in water, topsoil metals, etc.) and a given health outcome (e.g., cancer incidence, acute gastrointestinal infection hospital admissions, etc.) using available data. The review included all methodologies regardless of the aim of the environmental studies. For these reasons, we expected all or most of the articles considered in the review to be about cross-sectional studies. Focus of the review was to identify and describe materials, methods, and software programs. Therefore, the review does not present specific results.

2 Methods

In May 2019, we carried out a systematic literature search on articles describing ecological environmental methods using PubMed (National Center for Biotechnology Information 2019).

After a few tests to assess the most appropriate search terms to be used, we applied the following sequence of terms related with logical operators: "*(spatial analysis OR geographic analysis OR GIS) AND (water pollution OR water pollutants OR soil pollutants) AND (cancer registry OR population-based OR estimate OR estimating OR cancer incidence OR cancer mortality).*"

As a second step, we defined exclusion criteria, as follows: (a) articles on air pollutants not included in the project aims; (b) articles without real health outcome data such as risk assessment studies; (c) articles with ad-hoc data collection such as interviews or blood tests; (d) articles without spatial analysis; (e) articles not published in English.

The article revision process followed three phases. In *Phase 1,* three reviewers independently examined the abstracts of the articles identified by the PubMed search, so as to identify those potentially pertaining to our project aims. Articles would be considered eligible for Phase 2 if they were cleared by at least one of the reviewers. In *Phase 2*, four reviewers independently read the complete articles identified in *Phase 1*. In *Phase* 3, the reviewers met to address any divergence over *Phase 2* revisions.

At this stage, we then described the articles according to the following topics: country (or region) where the study was conducted; health outcome (dependent variable); environmental factors under analysis; socio-economic variables considered; smallest area unit considered for dependent variable; smallest area unit considered for environmental factor(s); final smallest area unit considered in the analysis; methods used; software used. Finally, we classified all selected papers into four subgroups, according to the methodology used and/or the data considered (a summary of characteristics is provided in Tables 1, 2, 3, 4, and 5 show articles by group).

Table 1 Synthesis of type of analysis to be performed for feasibility studies between water and soil pollutants and health outcomes, according to available data

	Environmental factor data	Health outcome data	Analysis	Number of articles
Type 1	Data by geographical areas	Data by geographical areas	Regression models using data by geographical areas	20
Type 2	Data at individual level	Data at individual level	Regression models using data at individual level	4
Type 3	Data by geographical areas	Data by geographical areas	Threshold values for exposure intensity are computed, in order to define cut-off points for evaluating trends in the health outcome variable influenced by the environmental factor	9
Type 4	Environmental pollution geographic clusters obtained by considering environmental factors and their potential emission sources	Clusters of areas or people generated by the analysis of the considered health outcomes	The two different kinds of clusters were identified separately. Comparisons between health outcomes geographic clusters and environmental pollution geographic clusters by considering the distance between them	7

Table 2 Articles classified as Type 1 and main characteristics

Reference PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Aballay et al. (2012); PMID: 22017596	Cancer incidence and pattern of arsenic concentration in drinking water wells in Cordoba, Argentina.	Cordoba province (Argentina)	5 cancer sites incidence	Arsenic in water	Gender, age, urban/rural residence	Districts	Sampling points in districts	Districts	Generalized linear latent and mixed model (GLLAMM). Likelihood ratio tests (LRT) were performed using the equivalent Poisson regression model for the random intercept model. Statistical significance at $p < 0.01$	STATA 10 with xtmepoisson command
Armijo and Coulson (1975); PMID: 23682416	Epidemiology of stomach cancer in Chile – The role of nitrogen fertilizers.	Chile	Mortality by stomach cancer	Nitrates in drinking water and nitrogen fertilizers	Infant mortality rates, housing ratings	Province	Province	Province	Bivariate correlation. Statistical significance at $p < 0.05$	Not declared
Bulka et al. (2016); PMID: 27136670	Arsenic in drinking water and prostate cancer in Illinois counties: An ecologic study.	Illinois state Cancer registry (USA)	Prostate cancer incidence	Arsenic (in drinking water)	Percent of individuals in the county living under the federal poverty level	County	County	County	Poisson regression model with robust standard errors. The model residuals were tested for spatial autocorrelation by calculating a global Moran's I statistic. Statistical significance at $p < 0.05$, $p < 0.01$	SAS 9.4

Chiang et al. (2010); PMID: 21139868	Spatiotemporal trends in oral cancer mortality and potential risks associated with heavy metal content in Taiwan soil.	Taiwan	Oral cancers age-standardized mortality rates	8 heavy metals (As, Cd, Cr, Cu, Hg, Ni, Pb, Zn) in soil	No	Townships	Townships	Townships	Factor analysis. Moran's I. SLM (spatial regression method, which can incorporate spatial dependency into the classical regression model). Monte Carlo estimation. Statistical significance at $p < 0.05$	GeoDa 0.9.5-I
Colak et al. (2015); PMID: 25619041	Geostatistical analysis of the relationship between heavy metals in drinking water and cancer incidence in residential areas in the Black Sea region of Turkey.	Black Sea region (Turkey)	Overall cancer incidence	17 heavy metal elements	No	Village/ district	Water sources inside the villages/districts	Village/ district	Kriging method. Linear regression analysis. Statistical significance at $p < 0.05$, $p < 0.01$	ArcGIS 10. SPSS 10
Hanchette et al. (2018); PMID: 30065203	Ovarian Cancer incidence in the U.S. and toxic emissions from pulp and paper plants: A geospatial analysis.	45 federal states and Washington D.C. (USA)	White females ovarian cancer incidence	Toxic air and water releases from pulp and paper mills	Only white females	County	ZIP code, county, and EPA region	County	Exploratory spatial data analysis: Moran's I and local indicator of spatial autocorrelation (LISA). Ordinary least squares (OLS) regression first for both the state- and county-level data. Spatial lag models for the state-level data. For the county-level data, GWR models.	ArcGIS 10.5; GeoDa

(continued)

Table 2 (continued)

Reference PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
									Statistical significance at $p < 0.05$, $p < 0.01$, $p < 0.001$	
Hendryx et al. (2012); PMID: 22471926	Permitted water pollution discharges and population cancer and non-cancer mortality: toxicity weights and upstream discharge effects in US rural–urban areas	Urban–rural areas (USA)	Mortality rates for cancer, kidney disease, and total non-cancer causes	Permitted toxic chemical pollutants in surface waters	College education rates, poverty rates, race/ethnicity percentages, rural–urban	County	County	County	Descriptive statistics and examination for multicollinearity, followed by non-spatial and spatial analyses (GWR). Statistical significance at $p < 0.01$	Not found
Huang et al. (2013); PMID: 23575356	Cell-type specificity of lung cancer associated with low-dose soil heavy metal contamination in Taiwan: an ecological study.	Taiwan	Lung cancer incidence	7 heavy metals (As, Cd, Cr, Cu, Hg, Pb, Ni, Zn) concentrations in soil	Sex, age.	Townships	Townships	Townships	Poisson regression models. Statistical significance at $p < 0.05$	SAS 9.13
Jian et al. (2017); PMID: 27713110	Associations between Environmental Quality and Mortality in the Contiguous United States, 2000–2005.	County by rural–urban continuum (USA)	All-cause mortality rate, heart disease, cancer, stroke	Environmental quality index (EQI)	Rural–urban continuum codes, percent of white population and the population density	County	County	County	Linear regression model to assess the average effects for the contiguous United States. Random intercept, random slope hierarchical model clustered	R 3.2.0 with the package lme4

									by different covariates. Statistical significance at $p < 0.05$	
Lin et al. (2014); PMID: 24566045	Assessing and mapping spatial associations among oral cancer mortality rates, concentrations of heavy metals in soil, and land use types based on multiple scale data.	Taiwan	Oral cancers age-standardized mortality rates	7 heavy metals (As, Cd, Cr, Cu, Pb, Ni, Zn) concentrations in soil	No	District	1 km × 1 km grid scale	1 km × 1 km grid scale	ATP Poisson kriging estimation. Anselin local Moran's I. Statistical significance at $p < 0.05$, $p < 0.001$	R
López-Abente et al. (2018a); PMID: 28155030	Compositional analysis of topsoil metals and its associations with cancer mortality using spatial misaligned data.	Spanish towns (Spain)	Mortality for 13 types of malignant tumors	Topsoil metal concentrations	Socio-demographic indicators: Population size, percentages of illiteracy, farmers, unemployment, average number of persons per household, mean income.	Town area (municipality)	Sampling locations	Cells 5 × 5 km	Kriging estimation. Factor analysis. BYM model with integrated nested Laplace approximations. Statistical significance at $p < 0.05$	R with the geoR, StatDA, and INLA packages
López-Abente et al. (2018b); PMID: 28847132	Residential radon and cancer mortality in Galicia, Spain.	Galicia (Spain)	14 cancer sites incidence	Radon/Arsenic (in topsoil)	Socio-demographic indicators: population size, percentages of illiteracy, farmers, unemployment, average number of persons per household, mean income.	Town area (municipality)	Sampling locations	Cells 10 × 10 km	BYM model with integrated nested Laplace approximations. Statistical significance at $p < 0.05$	R with the INLA package

(continued)

Table 2 (continued)

Reference PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Messier and Serre (2017); PMID: 27639278	Lung and stomach cancer associations with groundwater radon in North Carolina, USA.	North Carolina central Cancer registry (USA)	Stomach cancer and lung cancer incidence	Groundwater radon concentration (Bq/L)	Age, gender, race, residential tenure	Census tract	Census tract	Census tract	Negative binomial GLM with standard NB2 parameterization. Anselin local Moran's I. spatial autocorrelation of model residuals is assessed by examining a spatial covariance plot of the model standardized Pearson residuals. If present, a generalized estimating equation (GEE), which accounts for correlations between clusters and assumes no correlation within clusters, is implemented. Statistical significance at $p < 0.05$	R with the COUNT and GEE packages. BMElib numerical toolbox in MATLAB. Cluster and outlier analysis tool in ArcGIS 10.0.
Núñez et al. (2016); PMID: 27239676	Arsenic and chromium topsoil levels and cancer mortality in Spain.	Spanish towns (Spain)	Mortality for 27 types of malignant tumors	Arsenic and chromium (in topsoil)	Socio-demographic indicators: Population size, percentages of illiteracy, farmers, unemployment, average number of persons per household, mean income.	Town area (municipality)	Sampling locations	Town area (municipality)	Kriging estimation. Factor analysis. BYM model with integrated nested Laplace approximations. Statistical significance at $p < 0.05$	R with the INLA package

Núñez et al. (2017); PMID: 28108922	Association between heavy metal and metalloid levels in topsoil and cancer mortality in Spain.	Spanish towns (Spain)	Mortality for 27 types of malignant tumors	Topsoil metal concentrations	Socio-demographic indicators: Population size, percentages of illiteracy, farmers, unemployment, average number of persons per household, mean income.	Town area (municipality)	Sampling locations	Town area (municipality)	Kriging estimation. Factor analysis. BYM model with integrated nested Laplace approximations. Statistical significance at $p < 0.05$	R with the geoR, StatDA, and INLA packages
Ren et al. (2014); PMID: 25546281	Association between changing mortality of digestive tract cancers and water pollution: a case study in the Huai River Basin, China.	Huai River Basin (China)	Digestive cancer mortality	A series of frequency of serious pollution (FSP) indices including water quality grade (FSPWQG), biochemical oxygen demand (FSPBOD), chemical oxygen demand (FSPCOD), and ammonia nitrogen (FSPAN)	Gross domestic product	County	County	County	Linear correlation. Statistical significance at $p < 0.10$; $p < 0.05$, $p < 0.01$	Not declared
Roh et al. (2017); PMID: 28841521	Low-level arsenic exposure from drinking water is associated with prostate cancer in Iowa.	87 out of the 99 Iowa counties (USA)	White males prostate cancer incidence	Arsenic (in drinking water)	Poverty rate (only white males)	County	County	County	Spatial Poisson regression model. Anselin local Moran's I. Statistical significance at $p < 0.05$	SAS 9.4

(continued)

Table 2 (continued)

Reference PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Saint-Jacques et al. (2018); PMID: 29089168	Estimating the risk of bladder and kidney cancer from exposure to low-levels of arsenic in drinking water, Nova Scotia, Canada.	Nova Scotia (Canada)	Bladder cancer and kidney cancer incidence	Arsenic (in drinking water)	Area-based composite indices of material and social deprivation	Set of continuous 25 km^2 cells	Set of continuous 25 km^2 cells	Set of continuous 25 km^2 cells	BYM model with integrated nested Laplace approximations. Statistical significance at $p < 0.05$	R with the disease mapping and INLA packages
Su et al. (2010); PMID: 20152030	Incidence of oral cancer in relation to nickel and arsenic concentrations in farm soils of patients' residential areas in Taiwan.	Taiwan	Oral cancers age-standardized mortality rates	8 heavy metals (As, Cd, Cr, Cu, Hg, Ni, Pb, Zn) in soil	Personal income, factory density, factory distribution and types of industry, and other socio-economic variables	Township/ precinct	Township/ precinct	Township/ precinct	Step-wise multiple regression. Global Moran's I. Spatial models including conditional autoregressive model (CAR) and spatial simultaneous autoregressive (SAR) model. Statistical significance at $p < 0.05$	S-plus with spatial module
Van Leeuwen et al. (1999); PMID: 10597979	Associations between stomach cancer incidence and drinking water contamination with atrazine and nitrate in Ontario (Canada) agroecosystems, 1987–1991.	Ontario Cancer registry (Canada)	Age-standardized cancer incidence ratios: Stomach, colon, ovary, bladder, central nervous system, non-Hodgkin's lymphoma	Atrazine and nitrate in agroecosystems	Education level, income, occupation	Census sub-division (CSD)	Ecodistricts	Census sub-division (CSD)	Descriptive statistics and omnibus test. Least squares regression analysis. Global Moran's I. Statistical significance at $p < 0.25$, $p < 0.15$, $p < 0.05$	SPACESTAT

PMID PubMed identifier

Table 3 Articles classified as Type 2 and main characteristics

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Dahl et al. (2013); PMID: 22569744	Is the quality of drinking water a risk factor for self-reported forearm fractures? Cohort of Norway.	Norway	Forearm fractures	Main water quality indicators.	Marital status; education level; urban–rural residence	Geographic coordinates	Geographic coordinates	Geographic coordinates	GAM. Statistical significance at $p < 0.05$	ArcGIS 9.3. STATA 11
Edwards et al. (2014); PMID: 24506178	Regional specific groundwater arsenic levels and neuropsychological functioning: a cross-sectional study.	Texas Alzheimer's research and care consortium (USA)	TARCC neuropsychology scores	Arsenic in groundwater	Age, gender, education	Region	Cells of 0.8 square miles	Region	Linear regression models. Statistical significance at $p < 0.05$	ArcGIS
McDermott et al. (2014); PMID: 24771409	Does the metal content in soil around a pregnant woman's home increase the risk of low birth weight for her infant?	South Carolina (USA)	Low birth weight	8 heavy metals (As, Ba, Cr, Cu, Pb, Mn, Ni, Hg) in soil	Maternal age and race; number of priorbirths	GIS coordinates	GIS coordinates	GIS coordinates	Multivariable GAM. Statistical significance at $p < 0.001$	ArcGIS9.3. R with the mgcv package

(continued)

Table 3 (continued)

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Monrad et al. (2017); PMID: 28157645	Low-level arsenic in drinking water and risk of incident myocardial infarction: A cohort study.	Denmark	Myocardial infarction incidence	Arsenic in drinking water	Education level	Individual	Water supply area	Individual	Time-weighted average concentration. Evaluation of the exposure–response association by a cubic spline function with continuous first and second derivatives with 3 and 6 knots. Poisson GLM model. Statistical significance at $p < 0.05$	SAS (Lexis macro and PROC GENMOD procedure)

PMID PubMed identifier

Table 4 Articles classified as Type 3 and main characteristics

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Banning and Benfer (2017); PMID: 28820453	Drinking water uranium and potential health effects in the German Federal State of Bavaria.	Bavaria federal state (Germany)	Cancer and other diseases incidence	Uranium in drinking water	No	County	Municipality	Counties	Municipality concentration level used for the entire county, where available. Then classification in groups. Pearson correlation between SIR and concentration groups. Statistical significance at $p < 0.05$, $p < 0.01$	ArcGIS 10.1. SPSS
Cech et al. (1987); PMID: 3610447	Health significance of chlorination byproducts in drinking water: The Houston experience.	Houston, Texas (USA)	Mortality by urinary tract cancer, respiratory cancers, non-cancer respiratory causes	Trihalomethanes in drinking water	Gender, age, race	Census tract	Census tract	Census tract	Trends compared to pollutant concentration. Statistical significance at $p < 0.05$, $p < 0.01$	Not declared

(continued)

Table 4 (continued)

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Collman et al. (1988); PMID: 3198278	Radon-222 concentration in groundwater and cancer mortality in North Carolina.	North Carolina (USA)	Deaths from cancers of the nasal cavities, oro-, naso-, and hypopharynx, larynx, esophagus, stomach, colon, breast, bone, and the four major types of leukemia	Radon in public water supply	No	County	County	County	Relative risk by radon concentration. Statistical significance at $p < 0.05$	Not declared
Crump et al. (1987); PMID: 3591777	Cancer incidence patterns in the Denver metropolitan area in relation to the rocky flats plant.	Rocky flats, Colorado (USA)	Various cancers incidence	Plutonium in soil	Gender, age	Census tract	Census tract	Census tract	Bivariate analyses. Mantel-Haenszel test. Statistical significance at $p < 0.05$, $p < 0.01$, $p < 0.001$	Not declared
Dreiher et al. (2005); PMID: 16330453	Non-Hodgkin's lymphoma and residential proximity to toxic industrial waste in southern Israel.	Southern Israel	Non-Hodgkin's lymphoma incidence and survival	Toxic industrial waste	Gender, age, ethnicity, occupation	14-kms. Radius near the pollution source	14-kms. Radius near the pollution source	14-kms. Radius near the pollution source	GIS standardized rates. Kaplan–Meier method. Cox proportional hazard regression. Statistical significance at $p < 0.05$	MapInfo and not declared

Grilc et al. (2015); PMID: 27646727	Drinking water quality and the geospatial distribution of notified gastro-intestinal infections.	Slovenia	Acute gastrointestinal infections incidence	Fecal contamination of water supply system	No	Water supply zone	Water supply zone	Water supply zone	Classification of contaminated zones in three groups. Comparison with incidence in the same areas, computing the RRs. Statistical significance at $p < 0.05$	ArcGIS 10. Oracle 11 g
Richmond et al. (1987); PMID: 3616722	Colorectal cancer mortality and incidence in Campbell County, Kentucky.	Campbell County, Kentucky (USA)	Colon-rectum cancer incidence and mortality	Trihalomethanes in kitchen tap water	Gender, age, occupation	Census block	Census block	Census block	SIR and SMR compared to pollutant concentration. Statistical significance based on the Poisson distribution method of Bailar and Ederer. Statistical significance at $p < 0.05$	Not declared
Sánchez-Díaz et al. (2018); PMID: 30423874	Geographic analysis of motor neuron disease mortality and heavy metals released to Rivers in Spain	Spanish rivers	Deaths from motor neuron disease	Arsenic, cadmium, copper, chromium, mercury, lead, zinc in waters	No	Municipality	20 kms. of the rivers section from the emission point	Municipality	Log-linear models (Poisson link function). Statistical significance at $p < 0.05$, $p < 0.001$	Stata. ArcGIS

(continued)

Table 4 (continued)

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Thorpe and Shirmohammadi (2005); PMID: 16291529	Herbicides and nitrates in groundwater of Maryland and childhood cancers: a geographic information systems approach.	Maryland (USA)	4 childhood cancers incidence	Herbicides and nitrates	Gender, age, race	ZIP code	ZIP code	ZIP code	Cluster analysis. Contingency tables with chi-square analysis. Statistical significance at $p < 0.05$	ArcView 3.1. Spatial analyst 1.1. SaTScan 2.1. GraphPad prism 3.02

PMID PubMed identifier

Table 5 Articles classified as Type 4 and main characteristics

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Christian et al. (2011); PMID: 22043094	Exploring geographic variation in lung cancer incidence in Kentucky using a spatial scan statistic: Elevated risk in the Appalachian coal-mining region.	Kentucky (USA)	Lung cancer incidence	Coal mining waste and cigarette smoking	Gender, age	Circle areas	Circle areas	Circle areas	Discrete Poisson model. Monte Carlo simulation. Statistical significance at $p < 0.01$	SaTScan. ArcGIS 9.3
Cui et al. (2019); PMID: 30836673	Spatiotemporal variations in gastric Cancer mortality and their relations to influencing factors in S County, China	S County (China)	Gastric cancer mortality	Surface water quality	Population density, GDP	2x2 kms. Grid squares	2x2 kms. Grid squares	2x2 kms. Grid squares	Anselin local Moran's I. hot spot analysis. GeoDetector. Statistical significance at $p < 0.05$	ArcGIS 10.2. GeoDetector
Dai and Oyana (2008); PMID: 18939976	Spatial variations in the incidence of breast cancer and potential risks associated with soil dioxin	The Bay, Midland, and Saginaw counties, Central	Breast cancer incidence	Dioxin in soil	Age	ZIP code	ZIP code	ZIP code	Evaluation of soil dioxin contamination by using descriptive statistics and the SOM algorithm.	SOM toolbox. MatLab 7.1. ArcGIS 9.2. SatScan 7.0

(continued)

Table 5 (continued)

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
	contamination in Midland, Saginaw, and Bay Counties, Michigan, USA.	Michigan (USA)							Evaluation of the association between breast cancer rates and the ZIP codes by estimating the odds ratio and their corresponding 95% confidence intervals. Cluster detection using Kulldorff's spatial and space-time scan statistics and genetic algorithms for spatial and space-time clustering. Statistical significance at $p < 0.05$, $p < 0.01$, $p < 0.001$	

Fei et al. (2018); PMID: 29679198	The association between heavy metal soil pollution and stomach cancer: a case study in Hangzhou City, China	Hangzou city (China)	Stomach cancer incidence	Heavy metals in soil	Gender	Township	Sampling points	Township	Spatial distribution of incidence tested by Global Moran's I. GeoDetector. Hotspot analysis for environmental factor's cluster. Kriging's interpolation. Statistical significance at $p < 0.01$	GeoDetector
Guajardo and Oyana (2009); PMID: 20049167	A critical assessment of geographic clusters of breast and lung cancer incidences among residents living near the Tittabawassee and Saginaw Rivers, Michigan, USA.	The Bay, Midland, and Saginaw counties, Central Michigan (USA)	Breast and lung cancer incidences	Various pollutant and pollutants	Median household income, race, percent of native born, education level, percent of population residing at the same address in 1995	ZIP code	ZIP code	ZIP code	Preliminary GIS analysis. Odds ratio statistics. Stepwise discriminant function analysis. Ordinary Kriging. Anselin Local Moran's I. Turnbull's method. Bithell's linear risk score test. Lawson and Waller score test. Statistical significance at $p < 0.05$, $p < 0.01$	ArcGIS 9.2. ArcView 3.3. GeoDa 0.95i. ClusterSeer 2.0 and TerraSeer's STIS 1.6. Excel. SPSS 17.0

(continued)

Table 5 (continued)

Reference; PMID	Title	Country/ region	Dependent variable	Environmental factors	Socio-economic variables	Smallest area unit (dep. variable)	Smallest area unit (envir. factor)	Final smallest area unit considered	Methods	Software
Nieder et al. (2009); PMID: 19450849	Bladder cancer clusters in Florida: Identifying populations at risk.	Florida (USA)	Bladder cancer incidence	Arsenic in water	Race/ethnic categories, census derived poverty status at the block group level, census derived county-level urban/rural residence	Census block	Census block	Census block	Multivariate logistic regression. Statistical significance at $p < 0.05$, $p < 0.001$	ArcGIS 9.0. SaTScan 5.0. SPSS 11.0.1
Selvin et al. (1987); PMID: 3476785	Spatial distribution of disease: Three case studies.	Rocky flats, Colorado; Contra Costa County, California; Santa Clara County (California)	Lung cancer and leukemia incidence	Industrial facilities as proxy of pollution (pollutants not specified)	Gender, age, race	Census tract	Facility's position and distance from the cases	Census tract	Cluster analysis. Statistical significance at $p < 0.05$	Not declared

PMID PubMed identifier

3 Results

The PubMed search identified 694 articles. In *Phase 1* of the revision process, the reviewers agreed over 88% of the articles (considering both accepted and rejected articles). At this stage, 122 articles resulted eligible to be included in *Phase 2*. The complete read-through of 122 articles in *Phase 2* lead to an agreement of 61% among the four reviewers. After *Phase 3* of the revision, 40 articles were included in the review and classified as shown in Tables 2, 3, 4, and 5. Twenty of the articles referred to studies conducted in North America, 11 in Europe, 7 in Asia, and 2 in South America. The majority of articles (33 of 40) analyzed cancer incidence or mortality rates as outcome indicators. As for contaminants, 20 and 12 articles, respectively, analyzed pollutants in water and soil, while 7 articles analyzed pollutants in both elements and 1 article reported the results of applying the Environmental Quality Index to overall and by-cause mortality.

The statistical methods across the studies were quite diverse but could be grouped into specific families: descriptive analysis, data reduction procedures (factor analysis, cluster analysis), Moran's I and Kriging method for spatial interpolation, spatial regression analysis, various kinds of GLM regression models (often Poisson regression models), general additive models, Bayesian models with or without integrated nested Laplace approximations, and Monte Carlo estimations.

The authors of the studies we considered tested their results' statistical significance using different techniques; over half of the papers however did not report how they tested it (22). The remaining 18 articles used t-test (3 articles), chi-square, F-test, likelihood ratio test (2 articles for each test), Z score, Kruskal–Wallis test, Getis–Ord Gi statistic, Kulldorff's spatial and space-time scan statistics, Lawson and Waller score test, Mantel–Haenszel test and contingency table test, Poisson distribution method of Bailar and Ederer, Taylor series variance estimates, and a parametric bootstrap on testing for RR < 1.1 (1 article for each test). Statistical significance thresholds (p values) were reported in the Methods column of Tables 2, 3, 4, and 5.

As to packages, ArcGIS/ArcView and R are those principally used (14 articles each). See Fig. 1 for an overview of software use across the different studies.

The articles were synthetically classified in four groups, as shown in Table 1.

3.1 Type 1: Regression with Data by Geographical Area

Twenty articles (50%) were classified as belonging to this group (Table 2). Authors used different kinds of regression models to explore the relationship between health outcomes (dependent variable, e.g. cancer incidence or mortality), environmental factors, and any other covariate (e.g., socio-economic indicators) by geographical area. The geographic unit used to collect information on health outcome and pollution did not always match.

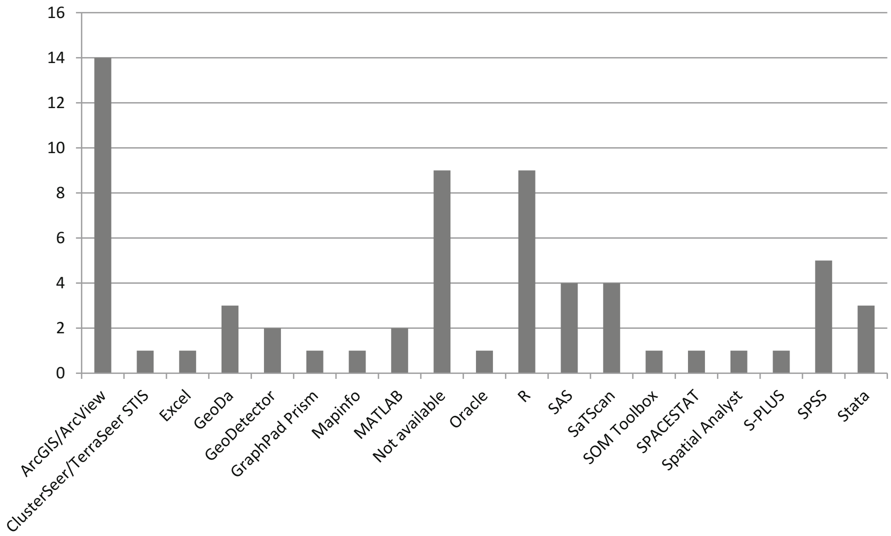

Fig. 1 Frequency of statistical software packages used

Eleven of the articles considered the same geographic areas for pollutants and for health outcomes (the areas either coincided or were very similar, e.g. districts, townships, etc.). In nine of the studies, information on pollution was collected using smaller geographic units than those used for health outcomes and addressed data misalignment in different ways. In five articles, authors applied the Kriging interpolation method (a Gaussian process regression) to extend the information collected at the pollution sources to the areas considered for epidemiologic population data (Colak et al. 2015; Lin et al. 2014; López-Abente et al. 2018a; Núñez et al. 2016, 2017). Authors of the remaining four articles used different approaches: the study by Hanchette et al. (2018) on the potential effects of toxic water releases on ovarian cancer incidence developed a combination of ordinary least squares (OLS) regression models and geographical weighted regression (GWR) models, corrected by spatial lag models (after testing the presence of local spatial autocorrelation by Local Indicators of Spatial Association – LISA – model). The study by López-Abente et al. (2018b) on the relationship between the presence of arsenic and radon in topsoil and 14 cancer sites incidence followed a Gaussian approach, which considered a Matérn Gaussian field approximated using the stochastic partial differential equation method for the environmental covariate. The study by Aballay et al. (2012) used a two-level model to estimate the effects of pollution in sampling points to the whole districts. Here, aquifer pollution was included as a random intercept and the misalignment was corrected by adaptive quadrature method. Finally, the study by Van Leeuwen et al. (1999) on atrazine and nitrate in drinking water and stomach cancer incidence determined mean contamination levels for each ecodistrict and data source; means were then proportionally combined to be associated to the population they represented.

Almost all 20 articles applied a preliminary exploratory spatial analysis using Moran's I for testing spatial autocorrelation after data geo-coding procedures.

As for regression models, the majority of studies incorporated different aspects of socio-economic, demographic, and life styles of the population under study in the evaluation of environmental exposure on the considered outcome. Only three articles (Chiang et al. 2010; Colak et al. 2015; Lin et al. 2014) did not correct the effects of such variables.

The choice of regression model varied between studies. Five studies applied the Besag, York, and Mollie's (BYM) model with integrated nested Laplace approximation. This is a convenient way to obtain approximations to the posterior marginal figures for parameters in Bayesian hierarchical models when the latent effects can be expressed as a Gaussian Markov random field, as it is defined in these works (López-Abente et al. 2018a, 2018b; Núñez et al. 2016, 2017; Saint-Jacques et al. 2018). Almost all other articles relied on different kinds of regression models, whether Poissonian or not, and considered the effects of spatial autocorrelation on the basis of the results of the Moran's I. Only the study by Armijo and Coulson (1975) on the relationship between stomach cancer mortality and presence of nitrate in drinking water and nitrogen fertilizers relied on bivariate correlation without considering the spatial autocorrelation term.

For the large part, two different kinds of software packages were used in these works, sometimes jointly, sometimes on their own. These were packages for geo-coding and study of spatial effects and model development.

R (S-Plus in one case) with its several packages was the most frequently used, because it made it easier to support geo-coding and analysis of spatial effects and to implement results in the Bayesian or regression models (Jian et al. 2017; Lin et al. 2014; López-Abente et al. 2018a, 2018b; Núñez et al. 2016, 2017; Messier and Serre 2017; Saint-Jacques et al. 2018; Su et al. 2010). Three studies relied on SAS for the versatility in adapting script adequate to combine the two aspects, as for R (Bulka et al. 2016; Huang et al. 2013; Roh et al. 2017). ArcGIS for spatial geo-coding and analysis was used in combination with other software packages (GeoDa and SPSS) in two studies (Colak et al. 2015; Hanchette et al. 2018); Stata, GeoDA, and Spacestat were used separately with few specific modules in 3 older studies (Aballay et al. 2012; Chiang et al. 2010; Van Leeuwen et al. 1999), while it was not possible to identify the software used for three of the articles (Armijo and Coulson 1975; Hendryx et al. 2012; Ren et al. 2014).

3.2 Type 2: Regression Models at Individual Level

Four articles (10%) pertained to this group (Table 3). The relationship between health condition/outcome and environmental factor was analyzed by regression models at individual level. The methodological interest was focused on the definition of an individual value for the environmental factor.

The geographic level for this group was mainly the individual, geo-coded at the geographic coordinates of residence (Dahl et al. 2013; McDermott et al. 2014; Monrad et al. 2017). Only Edwards et al. (2014) used the region of residence for attribution of exposure, but it mainly relied on the Texas Alzheimer's Research and Care Consortium (TARCC) neuropsychology scores at individual level for the analysis. Pollution sources were geo-coded at the same level (Dahl et al. 2013; McDermott et al. 2014) or at a slightly larger scale (water supply area (Monrad et al. 2017) or cells of 0.8 mile^2 (Edwards et al. 2014)). Pollution was then reported to the individual level by time-weighted average concentration and binary classification of exposure (McDermott et al. 2014, Monrad et al. 2017), stratification of exposure in groups (Dahl et al. 2013), or by attribution of the pollutant concentration in any cell to the corresponding person (Edwards et al. 2014).

All four studies used demographic and socio-economic covariates for correcting the environmental effects in different regression models such as linear regression (Edwards et al. 2014), Poisson generalized linear model (Monrad et al. 2017), and generalized additive model (Dahl et al. 2013, McDermott et al. 2014).

As with software packages, three articles used ArcGIS for data geo-coding (Dahl et al. 2013; Edwards et al. 2014; McDermott et al. 2014) and two of them combined other software packages for the models (R and Stata) (Dahl et al. 2013, McDermott et al. 2014). The study by Monrad et al. (2017) used SAS with specific procedures.

3.3 Type 3: Exposure Intensity Threshold Values for Evaluating Health Outcome

Nine articles (22.5%) were grouped as Type 3 (Table 4), characterized by a hiatus between the study of environmental factors and the distribution of health outcomes. The environmental factor, often detected as punctual source, was recoded as a categorical variable and the considered geographic areas were classified on the basis of the values/characteristics of such environmental categorical variable. The health outcome was analyzed at the same or larger area level. Therefore, threshold values for exposure intensity were computed in order to define cut-off points for evaluating trends in the health outcome variable to study the influence of the environmental factor.

In almost every article, the geographic areas considered for health outcomes, environmental factor, and other covariates were homogeneous; a few differences existed only in the studies by Banning and Benfer (2017) (county vs. municipality) and by Sánchez-Díaz et al. (2018) (municipality vs. river sections of 20 kms). In the first case, the pollutant concentration level in the municipality was extended to the entire county; in the second case, the case distance from the pollution source was considered as an independent variable of exposure in the final model.

Methods were not homogeneous due to differences in cut-off definitions and in the evaluation of their statistical significance with respect to the considered health outcome. Five articles also considered demographic and socio-economic characteristics of the population (Cech et al. 1987; Crump et al. 1987; Dreiher et al. 2005; Richmond et al. 1987; Thorpe and Shirmohammadi 2005).

ArcGIS and MapInfo were used for geo-coding, defining the cut-off points for the environmental factor and some spatial analysis (Banning and Benfer 2017; Dreiher et al. 2005; Grilc et al. 2015; Sánchez-Díaz et al. 2018; Thorpe and Shirmohammadi 2005); SPSS, SaTScan, and Stata for analyzing the potential correlation. In four articles, the used software packages were not declared (Cech et al. 1987; Collman et al. 1988; Crump et al. 1987; Richmond et al. 1987).

3.4 Type 4: Distance between Pollutant Source and Health Outcome Clusters

Seven articles (17.5%) belonged to this group (Table 5). No association between pollutants and health outcomes was considered in the first phase of these studies. Initially, they identified separately clusters of areas or people generated by the analysis of the considered health outcomes and environmental pollution geographic clusters obtained by considering environmental factors and their potential emission sources. As a second step, authors performed a comparison between health outcome geographic clusters and environmental pollution geographic clusters to evaluate their superimposition or proximity.

In five articles, the geographic areas considered for pollutants and health outcomes coincided (Christian et al. 2011; Cui et al. 2019; Dai and Oyana 2008; Guajardo and Oyana 2009; Nieder et al. 2009), thus reducing issues linked with the estimation of pollution concentration in areas wider than the one observed. The study by Selvin et al. (1987) on the relationship between leukemia, lung cancer incidence and industrial waste pollution used the distance between potential emission source and centroid of cases' residence census tract. This indicator became the factor connecting the cluster of disease with the cluster of pollution. The study by Fei et al. (2018) used the township of residence to geo-position the cases and a number of pollution sampling points in Hanghzou city; the authors joined this information with the hotspot analysis and the Kriging interpolation method so as to extend the pollutants concentration to the townships.

All articles used Moran's I as main indicator for evaluating spatial autoregression effects both on the environmental factors and the health outcomes. Also demographic, socio-economic, and life styles factors were considered in every work.

Different methods and techniques were used for the purpose of identifying environmental and health outcome clusters. These included classical cluster analysis (Selvin et al. 1987), Monte Carlo simulation and hypothesis testing for the identification of excess risk clusters (Christian et al. 2011; Nieder et al. 2009). Moreover, statistical analyses were performed by different score tests after combination of GIS and spatial techniques (Guajardo and Oyana 2009; Dai and Oyana 2008) and finally the quite recent GeoDetector, a spatial stratification statistical technique (Cui et al. 2019; Fei et al. 2018).

The studies in this group used a variety of software packages to address every specific issue, this is due to the peculiarity of these studies (all of them quite exploratory of not yet well-defined local situations). ArcGIS (in its various version) was used in almost every article for geo-coding and for some spatial analysis; SaTScan allowed to work in terms of “circles” of different, varying radius (Christian et al. 2011; Dai and Oyana 2008; Nieder et al. 2009); GeoDA, SPSS, ClusterSeer, TerraSeer and a few adaptable packages such as MATLAB, SOM Toolbox, and GeoDetector were used for finding and evaluating the statistical significance of the clusters (Cui et al. 2019; Dai and Oyana 2008; Fei et al. 2018; Guajardo and Oyana 2009).

4 Discussion

Our WASABY project herewith identifies and points out a number of public health studies that, regardless of their aims, may be of interest for the investigation of the relationship between environmental factors and health outcomes using available data.

The issue of exposure assessment has been investigated for over 40 years (the oldest study selected in this review dates 1975) and during this period significant changes were introduced in terms of the pollutants considered or in terms of the health outcomes analyzed. Innovations covered new technologies to measure pollutants, statistical methodologies to assess exposure, and software and hardware progress. These changes allowed to develop and apply geo-coding and statistical methods for the reduction of the ecological bias when considering the relationships among individuals, geographic areas, pollutants, and health outcomes (Woods et al. 2005).

More complex models for interpolation and analysis have become available with the development of software and hardware allowing for increased computation power. Most of the studies we considered were developed after the first decade of the twenty-first century (29 studies were published after 2009) when tools for spatial analysis and representation were greatly developed and made more user-friendly, thanks to the introduction of more powerful processors. This was particularly true for spatial interpolation and estimation of multifactor effects which used to be applied to large datasets. As an example, the Intel Core microprocessors (I3-I7) became available in 2010 offering superior computational power.

Following the growing demand for these types of studies, new packages and user-interfaces for free programs (e.g., R) and scripts for commercial programs (e.g., SAS, Stata) were developed. Specifically, procedures such as the Kriging interpolation, the computation of Moran’s I, the application of Poisson linear regression, or INLA models became more accessible after the introduction of new tools and improvements. Geographic representation programs markedly improved including internal tools for simple and more specific statistic analysis as well as more user-friendly interfaces thus widening the audience of users.

Criteria for software choice naturally include availability of specific tools/scripts for a) management and linkage of large datasets, b) spatial interpolation and advanced analysis (like the INLA models in R), c) geographic representation, and d) for cost. In consideration of the above-mentioned criteria, R is often considered the best choice thanks to the extension of available tools that allow to develop all procedures for free.

Commercial programs such as Stata and SAS offer more user-friendly interfaces at higher costs. For this reason, they are chosen by virtue of the availability and quality of the scripts.

As to geographic representations, ArcGIS (commercial), QGIS, and SaTScan (free) appear to be the best choice, owing to their usability, connection with online map sources, and presence of internal tools for both simple and more sophisticated spatial analyses.

A major merit of our study is the identification and critical evaluation of published articles on the topic by four individual researchers under standardized criteria and methods. In our review we highlighted some of the most recent studies, methodologies, and techniques able to define the smallest available units of observation (e.g., the census tract or specific small territorial cells defined in each research). This improved the estimation reliability of the effects on health due to the exposure to pollutants and other factors when transferring considerations from "area" to "person" (Lillini and Vercelli 2019), in compliance with EU privacy legislation on analyses at the individual level.

Our work does not intend to offer a comprehensive overview of methodologies for ecological environmental studies on water and soil pollution in relation to public health, as relevant articles on these issues might have been missed out as a result of the term search criteria. However, we hope to have intercepted most of the main relevant methodologies and techniques.

Another limitation of our study is the exclusion of non-English language papers. A number of articles written in Chinese, Italian, Russian, and Spanish were not considered in this work due to sub-optimal readability (Chinese and Russian ones) and comprehensibility (Spanish ones) as well as to enhance the possibility of reaching a wide audience.

The methods reported in this review are appropriate for research on water and soil pollution data, as detailed in the rationale of the WASABY project; for this reason, they could not be generalized to other environmental risk factors, such as air pollutants.

Finally, most of the considered works shared the cross-sectional study design, as expected.

Overall, our analysis shows a wide variation of valid and reliable methods and techniques. It is not possible to identify a "gold standard" because of the peculiarity of every situation. On the other hand, when approaching such issues, scholars may identify the research experiences that best fit the situation they are approaching to investigate, apply all corresponding procedures, and adapt them to the specific situation they are facing.

Here, we wish to give our insight on the use of different statistical models so as to provide some advice for choosing the best option for different research aims. First, researchers will have to choose whether to opt for a frequentist or Bayesian approach. This choice is both theoretical and practical (Samaniego 2010).

The frequentist approach assumes one's measurements are enough to state something meaningful. Probability is defined in terms of limiting frequency of occurrence of an event, it assumes that there are true values of the model parameters and it computes the parameters point estimates. In the Bayesian approach, data are supplemented with additional information in the form of a prior probability distribution. The prior belief about the parameters is combined with the data's likelihood function according to Bayes theorem, in order to yield the posterior belief about the parameters. Probability is the degree of belief on the occurrence of an event, only data are real and there are no true values of parameters as such, apart from the fact that a number of values are more probable than others.

Most frequently used models are linear ones, e.g., Poisson regression or general additive models (GAMs), Besag York Mollié (BYM) models with or without integrated nested Laplace approximation (INLA).

Poisson regression seems appropriate when the dependent variable is a count, the events must be independent, but the probability per unit time of events is understood to be related to covariates. Poisson regression is also appropriate for rate data, where the rate is a count of events divided by part of a given unit's exposure (a particular unit of observation). Event rates may be calculated as events per unit time, which allows the observation window to vary for each unit. Here, exposure is respectively unit area, person−years, and unit time (Tutz 2011).

GAMs are a class of statistical models in which the usual linear relationship between the response and predictors is replaced by several non-linear smooth functions to model and capture the non-linearity of data. These are flexible techniques that help to fit linear models which can either be linearly or non-linearly dependant on several predictors. The latter characteristic makes them very useful and reliable to identify and describe non-linear relationships between response and predictors. There are at least three good reasons for using GAM: interpretability, flexibility/automation, and regularization. When the model contains non-linear effects, GAM provides a regularized and interpretable solution, while other methods generally lack at least one of these three features (Hastie and Tibshirani 1990).

BYM model is a Bayesian hierarchical model based on a conditional autoregressive (CAR) model for spatial random effects. In the CAR model, spatial dependence is expressed conditionally: given the values in all other areas, it requires that the random effect in an area depends only on a small set of neighboring values. An essential aspect of the BYM model and its extensions is the specification of the neighborhood structure for the areas. This is quite flexible and it may be arbitrarily defined. It is based on adjacency relationships of the geographical areas (or disjoint geographical areas with the needed correction) (Rodrigues and Assunção 2012). BYM is useful to investigate the underlying relative risks of a disease observed on joint or disjoint geographical areas. On the other end, however, it needs a stable and quite homogeneous definition of the geographical units and outcomes, and covariates must be defined at the same geographical level or they should be interpolated at such level.

In some studies, BYM models were developed along with INLA, which relies on a combination of analytical approximations and efficient numerical integration schemes to achieve highly accurate deterministic approximations to posterior quantities of interest. The main benefit for using INLA instead of Markov chain Monte Carlo (MCMC) techniques is computation. INLA is fast even for large and complex models. Moreover, INLA is a deterministic algorithm and does not suffer from slow convergence and poor mixing (Rue et al. 2009).

A common aspects considered in spatial analysis is the spatial autocorrelation, i.e. the co-variation of properties within geographic space. Characteristics at proximal locations appear to be correlated, either positively or negatively. The spatial autocorrelation problem violates the condition of standard statistical techniques that assume independence among observations (Knegt et al. 2010). Spatial analysis models correct spatial autocorrelation with different techniques. In the studies analyzed by this review, spatial autocorrelation is measured by Moran's I, a correlation coefficient that measures the overall spatial autocorrelation of the data set. In other words, it measures how one object is similar to others surrounding it. If objects are attracted (or repelled) by each other, it means that the observations are not independent. The standardized version of Moran's I enables to compare the significant spatial patterns of different or same variables with different calculating parameters and it should be chosen as the preferred test (Getis 2010).

Another observation regards the convergence of the geographical level at which the data is collected. In several cases, health outcomes, environmental variables, and other covariates (e.g., socio-economic data) are collected at the same geographic level (e.g., municipality, census tract, etc.). In other cases areas do not coincide due to the different availability and data characteristics in the selected sources. When this is the case, interpolating methods must be applied to reduce territorial bias.

Many of the articles considered in our study used Kriging regression as the preferred method of interpolation so values are modeled by a Gaussian process governed by prior covariances. Under suitable assumptions on the priors, Kriging gives the best linear unbiased prediction of the intermediate values. Interpolating methods based on other criteria such as smoothness (e.g., smoothing spline) may not yield the most likely intermediate values. There are two Kriging methods: ordinary and universal. Ordinary Kriging is the most general and widely used of the Kriging methods. Here, the constant mean is assumed as unknown. Universal Kriging assumes that there is an overriding trend in the data which can be modeled by a deterministic function, a polynomial. Universal Kriging should only be used when you know there is a trend in your data and you can give a scientific justification to describe it. The method can be used where spatially-related data has been collected and estimates of "fill-in" data are desired in the locations (spatial gaps) between the actual measurements (Oliver and Webster 1990).

Another example of the research choice to be made in these types of studies is the use of socio-economic information as a correction of the effects of the environmental conditions on health outcomes. This correction should always be considered in such type of studies if the socio-economic data are available and reliable. This is because there is a relevant relationship between these characteristics, the probability of living in areas where exposure to pollution is significant, and the health condition of the considered population (see, for instance, the founding work of Dolk et al. 1995, or the more recent Pannullo et al. 2016). When socio-economic characteristics are not considered, the bias is a more superficial description of the interested population and it is possible to lose relevant indication to better address health policy actions. It is, therefore, advisable to collect socio-economic data, at least at small geographic area (e.g., Census Tract, Woods et al. 2005). In many countries, ecological socio-economic deprivation indices at such geographic level are already available (e.g., Guillaume et al. 2016; Caranci et al. 2010).

5 Conclusion

This review represents a useful tool for cancer registries, health institutions, and environmental agencies that are interested in territorial monitoring, health or environmental surveillance. We provide suggestions on methods, techniques, and tools which may be applied in studies that investigate disease clusters and environmental exposure. In this perspective, the study contributes to optimize the use of public health resources.

Acknowledgments This document was produced by "WASABY – Water And Soil contamination and Awareness on Breast cancer risk in Young women", which received funding from the European Commission as part of the framework of the Health Program (Contract No. PP-2-2016 with the DG SANTE Directorate C). *Wasaby Working Group*: FRANCE. *Registre général des cancers de Poitou-Charentes:* Gautier Defossez. *Registre des tumeurs de Loire-Atlantique et Vendée:* Florence Molinie; Anne Cowpli-Bony. *Registre général des cancers de Haute Vienne:* Tania Dalmeida. *Registre général des tumeurs du Calvados:* Anne Valérie Guizard. *Registre des cancers de la Manche:* Simona Bara. *Registre du cancer de la Somme:* Bénédicte Lapotre-Ledoux. *Registre général des cancers de Lille et de sa région:* Sandrine Plouvier. *Registre des cancers du Bas-Rhin:* Michel Velten. *Registre des cancers du Haut-Rhin:* Séverine Boyer. *Registre des tumeurs du Doubs:* Anne Sophie Woronoff. *Registre des cancers du sein et des cancers gynécologiques de Côte-d'Or:* Patrick Arveux. *Registre du cancer de l'Isère:* Marc Colonna. *Registre des tumeurs de l'Hérault:* Brigitte Tretarre. *Registre des cancers du Tarn:* Pascale Grosclaude. *Registre général des cancers de Guadeloupe:* Jacqueline Deloumeaux. *Registre général des cancers de Martinique:* Clarisse Joachim. *Registre des cancers de Gironde:* Gaelle Coureau. *Université de Caen:* Josephine Bryere ITALY. *Analytical Epidemiology and Health Impact Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milano:* Simone Bonfarnuzzo, Milena Sant. *Registro Tumori dell'Alto Adige:* Guido Mazzoleni; Fabio Vittadello. *Registro Tumori ASL Napoli 3 Sud:* Mario Fusco; Maria Francesca Vitale; Valerio Ciullo. *Registro Tumori di Palermo e provincia:* Rosanna Cusimano; Walter Mazzucco; Maurizio Zarcone. *Registro Tumori di Parma:* Maria Michiara; Paolo Sgargi. *Registro Tumori ASP Ragusa-Caltanisetta:* Carmela Nicita; Rosario Tumino. *Registro Tumori della Provincia di Siracusa:* Ylenia Dinaro; Francesco Tisano. *Registro Tumori di Trapani e*

Agrigento: Giuseppa Candela; Tiziana Scuderi. *Registro Tumori di Trento:* Roberto Rizzello; Silvano Piffer. *Registro Tumori Umbria:* Fabrizio Stracci; Fortunato Bianconi. *Registro Tumori di Varese:* Giovanna Tagliabue. LITHUANIA. *Lithuanian Cancer Registry:* Ieva Vincerzevskiene. POLAND. *Polish National Cancer Registry:* Joanna Didkowska; Urszula Wojciechowska; Krzysztof Czaderny. *Greater Poland Cancer Registry:* Łukasz Taraszkiewicz; Maciej Trojanowski. *Kielce Cancer Registry:* Pawel Macek. *Masovia Cancer Registry:* Urszula Sulkowska. *Silesia Cancer Registry:* Marcin Motnyk. *Subcarpatian Cancer Registry:* Monika Gradalska-Lampart. PORTUGAL. *Registo Oncológico Regional do Norte:* Luis Antunes; Jéssica Rodrigues. *Registo Oncológico Regional - Zona Centro:* Joana Antunes Lima Bastos; Margarida Ornelas. SPAIN. *Registro de Cáncer de Euskadi-CIBERESP:* Arantza Lopez de Munain; Nerea Larrañaga. *Registro de Tumores de Castellón-Valencia:* Paloma Botella; Consol Sabater Gregori. *Registro de Cáncer de Girona:* Marc Saez; Rafael Marcos-Gragera. *Registro de Cáncer de Granada, EASP, CIBERESP, ibs.GRANADA, UGR:* Maria Jose Sanchez-Perez; Miguel Rodriguez-Barranco. *Registro de Cáncer de Murcia-CIBERESP:* Monica Ballesta-Ruiz; Maria Dolores Chirlaque. *Registro de Cáncer de Navarra-CIBERESP:* Eva Ardanaz; Marcela Guevara. NORTHERN IRELAND (UK). *Northern Ireland Cancer Registry:* Anna Gavin; David Donnelly.

Declaration of Competing Financial Interests (CFI) The authors declare they have no actual or potential competing financial interests.

References

Aballay LR, Díaz Mdel P, Francisca FM, Muñoz SE (2012) Cancer incidence and pattern of arsenic concentration in drinking water wells in Córdoba, Argentina. Int J Environ Health Res 22 (3):220–231. https://doi.org/10.1080/09603123.2011.628792

Armijo R, Coulson AH (1975) Epidemiology of stomach cancer in Chile--the role of nitrogen fertilizers. Int J Epidemiol 4(4):301–309. https://doi.org/10.1093/ije/4.4.301

Baker D, Kjellström T, Calderon R, Pastides H (1999) Environmental epidemiology: a textbook on study methods and public health applications. Preliminary edition. World Health Organization, Malta

Banning A, Benfer M (2017) Drinking water uranium and potential health effects in the German Federal State of Bavaria. Int J Environ Res Public Health 14(8):E927. https://doi.org/10.3390/ijerph14080927

Bulka CM, Jones RM, Turyk ME, Stayner LT, Argos M (2016) Arsenic in drinking water and prostate cancer in Illinois counties: an ecologic study. Environ Res 148:450–456. https://doi.org/10.1016/j.envres.2016.04.030

Caranci N, Biggeri A, Grisotto L, Pacelli B, Spadea T, Costa G (2010) The Italian deprivation index at census block level: definition, description and association with general mortality. Epidemiol Prev 34(4):167–176

Cech I, Holguin AH, Littell AS, Henry JP, O'Connell J (1987) Health significance of chlorination byproducts in drinking water: the Houston experience. Int J Epidemiol 16(2):198–207. https://doi.org/10.1093/ije/16.2.198

Chiang CT, Lian IB, Su CC, Tsai KY, Lin YP, Chang TK (2010) Spatiotemporal trends in oral cancer mortality and potential risks associated with heavy metal content in Taiwan soil. Int J Environ Res Public Health 7(11):3916–3928. https://doi.org/10.3390/ijerph7113916

Christian WJ, Huang B, Rinehart J, Hopenhayn C (2011) Exploring geographic variation in lung cancer incidence in Kentucky using a spatial scan statistic: elevated risk in the Appalachian coal-mining region. Public Health Rep 126(6):789–796. https://doi.org/10.1177/003335491112600604

Colak EH, Yomralioglu T, Nisanci R, Yildirim V, Duran C (2015) Geostatistical analysis of the relationship between heavy metals in drinking water and cancer incidence in residential areas in the Black Sea region of Turkey. J Environ Health 77(6):86–93

Collman GW, Loomis DP, Sandler DP (1988) Radon-222 concentration in groundwater and cancer mortality in North Carolina. Int Arch Occup Environ Health 61(1–2):13–18

Crump KS, Ng TH, Cuddihy RG (1987) Cancer incidence patterns in the Denver metropolitan area in relation to the rocky flats plant. Am J Epidemiol 126(1):127–135. https://doi.org/10.1093/oxfordjournals.aje.a114644

Cui C, Wang B, Ren H, Wang Z (2019) Spatiotemporal variations in gastric cancer mortality and their relations to influencing factors in S County, China. Int J Environ Res Public Health 16(5): E784. https://doi.org/10.3390/ijerph16050784

Dahl C, Søgaard AJ, Tell GS, Flaten TP, Krogh T, Aamodt G, NOREPOS Core Research Group (2013) Is the quality of drinking water a risk factor for self-reported forearm fractures? Cohort of Norway. Osteoporos Int 24(2):541–551. https://doi.org/10.1007/s00198-012-1989-7

Dai D, Oyana TJ (2008) Spatial variations in the incidence of breast cancer and potential risks associated with soil dioxin contamination in Midland, Saginaw, and Bay Counties, Michigan, USA. Environ Health 7:49. https://doi.org/10.1186/1476-069X-7-49

Deener KCK, Sacks JD, Kirrane EF, Glenn BS, Gwinn MR, Bateson TF, Burke TA (2018) Epidemiology: a foundation of environmental decision making. J Expo Sci Environ Epidemiol 28(6):515–521. https://doi.org/10.1038/s41370-018-0059-4

Dolk H, Mertens B, Kleinschmidt I, Walls P, Shaddick G, Elliott P (1995) A standardisation approach to the control of socioeconomic confounding in small area studies of environment and health. J Epidemiol Community Health 49(Suppl 2):S9–S14. https://doi.org/10.1136/jech.49.suppl_2.s9

Dreiher J, Novack V, Barachana M, Yerushalmi R, Lugassy G, Shpilberg O (2005) Non-Hodgkin's lymphoma and residential proximity to toxic industrial waste in southern Israel. Haematologica 90(12):1709–1710

Edwards M, Johnson L, Mauer C, Barber R, Hall J, O'Bryant S (2014) Regional specific groundwater arsenic levels and neuropsychological functioning: a cross-sectional study. Int J Environ Health Res 24(6):546–557. https://doi.org/10.1080/09603123.2014.883591

European Commission (2016) Call for proposals for a pilot project on primary prevention courses for girls living in areas with higher risk of breast cancer. http://ec.europa.eu/research/participants/data/ref/other_eu_prog/other/hp/call-fiche/hp-call-fiche-pp2-5_en.pdf. Accessed 5 Aug 2019

Fei X, Lou Z, Christakos G, Ren Z, Liu Q, Lv X (2018) The association between heavy metal soil pollution and stomach cancer: a case study in Hangzhou City, China. Environ Geochem Health 40(6):2481–2490. https://doi.org/10.1007/s10653-018-0113-0

Getis A (2010) The analysis of spatial association by use of distance statistics. Geogr Anal 24 (3):189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261

Grilc E, Gale I, Veršič A, Žagar T, Sočan M (2015) Drinking water quality and the geospatial distribution of notified gastro-intestinal infections. Zdr Varst 54(3):194–203. https://doi.org/10.1515/sjph-2015-0028

Guajardo OA, Oyana TJ (2009) A critical assessment of geographic clusters of breast and lung cancer incidences among residents living near the Tittabawassee and Saginaw Rivers, Michigan, USA. J Environ Public Health 2009:316249. https://doi.org/10.1155/2009/316249

Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M, Marí-Dell'Olmo M, Fernández Fontelo A, Borrell C, Ribeiro AI, Pina MF, Mayer A, Delpierre C, Rachet B, Launoy G (2016) Development of a cross-cultural deprivation index in five European countries. J Epidemiol Community Health 70(5):493–499. https://doi.org/10.1136/jech-2015-205729

Hanchette C, Zhang CH, Schwartz GG (2018) Ovarian cancer incidence in the U.S. and toxic emissions from pulp and paper plants: a geospatial analysis. Int J Environ Res Public Health 15 (8). https://doi.org/10.3390/ijerph15081619

Hänninen O, Knol AB, Jantunen M, Lim TA, Conrad A, Rappolder M, Carrer P, Fanetti AC, Kim R, Buekers J, Torfs R, Iavarone I, Classen T, Hornberg C, Mekel OC, EBoDE Working

Group (2014) Environmental burden of disease in Europe: assessing nine risk factors in six countries. Environ Health Perspect 122(5):439–446. https://doi.org/10.1289/ehp.1206154
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, New York
Hendryx M, Conley J, Fedorko E, Luo J, Armistead M (2012) Permitted water pollution discharges and population cancer and non-cancer mortality: toxicity weights and upstream discharge effects in US rural-urban areas. Int J Health Geogr 11:9. https://doi.org/10.1186/1476-072X-11-9
Huang HH, Huang JY, Lung CC, Wu CL, Ho CC, Sun YH, Ko PC, Su SY, Chen SC, Liaw YP (2013) Cell-type specificity of lung cancer associated with low-dose soil heavy metal contamination in Taiwan: an ecological study. BMC Public Health 13:330. https://doi.org/10.1186/1471-2458-13-330
Jian Y, Messer LC, Jagai JS, Rappazzo KM, Gray CL, Grabich SC, Lobdell DT (2017) Associations between environmental quality and mortality in the contiguous United States, 2000-2005. Environ Health Perspect 125(3):355–362. https://doi.org/10.1289/EHP119
Knegt DE, Coughenour MB, Skidmore AK, Heitkönig IMA, Knox NM, Slotow R, Prins HHT (2010) Spatial autocorrelation and the scaling of species–environment relationships. Ecology 91 (8):2455–2465. https://doi.org/10.1890/09-1359.1
Lillini R, Vercelli M (2019) The local socio-economic health deprivation index: methods and results. J Prev med Hyg 59(4 Suppl 2):E3–E10. https://doi.org/10.15167/2421-4248/jpmh2018.59.4s2.1170
Lin WC, Lin YP, Wang YC, Chang TK, Chiang LC (2014) Assessing and mapping spatial associations among oral cancer mortality rates, concentrations of heavy metals in soil, and land use types based on multiple scale data. Int J Environ Res Public Health 11(2):2148–2168. https://doi.org/10.3390/ijerph110202148
López-Abente G, Locutura-Rupérez J, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Núñez O (2018a) Compositional analysis of topsoil metals and its associations with cancer mortality using spatial misaligned data. Environ Geochem Health 40(1):283–294. https://doi.org/10.1007/s10653-016-9904-3
López-Abente G, Núñez O, Fernández-Navarro P, Barros-Dios JM, Martín-Méndez I, Bel-Lan A, Locutura J, Quindós L, Sainz C, Ruano-Ravina A (2018b) Residential radon and cancer mortality in Galicia, Spain. Sci Total Environ 610-611:1125–1132. https://doi.org/10.1016/j.scitotenv.2017.08.144
McDermott S, Bao W, Aelion CM, Cai B, Lawson AB (2014) Does the metal content in soil around a pregnant woman's home increase the risk of low birth weight for her infant? Environ Geochem Health 36(6):1191–1197. https://doi.org/10.1007/s10653-014-9617-4
Messier KP, Serre ML (2017) Lung and stomach cancer associations with groundwater radon in North Carolina, USA. Int J Epidemiol 46(2):676–685. https://doi.org/10.1093/ije/dyw128
Monrad M, Ersbøll AK, Sørensen M, Baastrup R, Hansen B, Gammelmark A, Tjønneland A, Overvad K, Raaschou-Nielsen O (2017) Low-level arsenic in drinking water and risk of incident myocardial infarction: a cohort study. Environ Res 154:318–324. https://doi.org/10.1016/j.envres.2017.01.028
National Center for Biotechnology Information. U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pubmed. Accessed 5 Aug 2019
Nieder AM, MacKinnon JA, Fleming LE, Kearney G, Hu JJ, Sherman RL, Huang Y, Lee DJ (2009) Bladder cancer clusters in Florida: identifying populations at risk. J Urol 182(1):46–50. https://doi.org/10.1016/j.juro.2009.02.149
Núñez O, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Locutura JF, López-Abente G (2016) Arsenic and chromium topsoil levels and cancer mortality in Spain. Environ Sci Pollut Res Int 23(17):17664–17675. https://doi.org/10.1007/s11356-016-6806-y
Núñez O, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Locutura Rupérez JF, López-Abente G (2017) Association between heavy metal and metalloid levels in topsoil and cancer mortality in Spain. Environ Sci Pollut Res Int 24(8):7413–7421. https://doi.org/10.1007/s11356-017-8418-6

Oliver MA, Webster R (1990) Kriging: a method of interpolation for geographical information systems. Int J Geograph Inf Syst 4(3):313–332. https://doi.org/10.1080/02693799008941549
Pannullo F, Lee D, Waclawski E, Leyland AH (2016) How robust are the estimated effects of air pollution on health? Accounting for model uncertainty using Bayesian model averaging. Spat Spatiotemporal Epidemiol 18:53–62. https://doi.org/10.1016/j.sste.2016.04.001
Ren H, Wan X, Yang F, Shi X, Xu J, Zhuang D, Yang G (2014) Association between changing mortality of digestive tract cancers and water pollution: a case study in the Huai River Basin, China. Int J Environ Res Public Health 12(1):214–226. https://doi.org/10.3390/ijerph120100214
Richmond RE, Rickabaugh J, Huffman J, Epperly N (1987) Colorectal cancer mortality and incidence in Campbell County, Kentucky. South Med J 80(8):953–957. https://doi.org/10.1097/00007611-198708000-00005
Rodrigues EC, Assunção R (2012) Bayesian spatial models with a mixture neighborhood structure. J Multivar Anal 109:88–102. https://doi.org/10.1016/j.jmva.2012.02.017
Roh T, Lynch CF, Weyer P, Wang K, Kelly KM, Ludewig G (2017) Low-level arsenic exposure from drinking water is associated with prostate cancer in Iowa. Environ Res 159:338–343. https://doi.org/10.1016/j.envres.2017.08.026
Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B Stat Methodol 71:319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x
Saint-Jacques N, Brown P, Nauta L, Boxall J, Parker L, Dummer TJB (2018) Estimating the risk of bladder and kidney cancer from exposure to low-levels of arsenic in drinking water, Nova Scotia, Canada. Environ Int 110:95–104. https://doi.org/10.1016/j.envint.2017.10.014
Samaniego FJ (2010) A comparison of the Bayesian and frequentist approaches to estimation. Springer, New York. https://doi.org/10.1007/978-1-4419-5941-6
Sánchez-Díaz G, Escobar F, Badland H, Arias-Merino G, Posada de la Paz M, Alonso-Ferreira V (2018) Geographic analysis of motor neuron disease mortality and heavy metals released to Rivers in Spain. Int J Environ Res Public Health 15(11):E2522. https://doi.org/10.3390/ijerph15112522
Selvin S, Shaw G, Schulman J, Merrill DW (1987) Spatial distribution of disease: three case studies. J Natl Cancer Inst 79(3):417–423
Stewart BW, Wild CP (2014) World Cancer report 2014. International Agency for Research on Cancer, Lyon
Su CC, Lin YY, Chang TK, Chiang CT, Chung JA, Hsu YY, Lian IB (2010) Incidence of oral cancer in relation to nickel and arsenic concentrations in farm soils of patients' residential areas in Taiwan. BMC Public Health 10:67. https://doi.org/10.1186/1471-2458-10-67
Thorpe N, Shirmohammadi A (2005) Herbicides and nitrates in groundwater of Maryland and childhood cancers: a geographic information systems approach. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 23(2):261–278. https://doi.org/10.1080/10590500500235001
Tutz G (2011) Poisson regression. In: Lovric M (ed) International encyclopedia of statistical science. Springer, Berlin. https://doi.org/10.1007/978-3-642-04898-2_450
Van Leeuwen JA, Waltner-Toews D, Abernathy T, Smit B, Shoukri M (1999) Associations between stomach cancer incidence and drinking water contamination with atrazine and nitrate in Ontario (Canada) agroecosystems, 1987-1991. Int J Epidemiol 28(5):836–840. https://doi.org/10.1093/ije/28.5.836
Wasaby Website. http://www.wasabysite.it/. Accessed 5 Aug 2019
Woods LM, Rachet B, Coleman MP (2005) Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. Br J Cancer 92(7):1279–1282. https://doi.org/10.1038/sj.bjc.6602506
World Health Organization Website. Health policy page. http://www.euro.who.int/en/health-topics/health-policy. Accessed 5 Aug 2019