ENVIRONMENTAL RISK ASSESSMENT IN THE MEDITERRANEAN REGION USING ARTIFICIAL NEURAL NETWORKS

Marelys Josefina Mújica Chacín

Dipòsit Legal: T. 1057-2012

UNIVERSITAT ROVIRA I VIRGILI

Marelys Josefina Mujica Chacín

ENVIRONMENTAL RISK ASSESSMENT IN THE MEDITERRANEAN REGION USING ARTIFICIAL NEURAL NETWORKS

DOCTORAL DISSERTATION

Tarragona, Spain
2012

Marelys Josefina Mujica Chacín

ENVIRONMENTAL RISK ASSESSMENT IN THE MEDITERRANEAN REGION USING ARTIFICIAL NEURAL NETWORKS

DOCTORAL DISSERTATION

Directed by Dr. Francesc Giralt and Dr. Robert Rallo

Departament d'Enginyeria Química

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2012

UNIVERSITAT ROVIRA I VIRGILI

Universitat Rovira i Virgili
Departament d'Enginyeria Química

Campus Sescelasdes, Av. Països Catalans 26,
46007 Tarragona, Spain
Tel: 977559700
Fax: 977559699

We, Dr. Francesc Giralt I Prat and Dr. Robert Rallo Moya, members of the Department of Chemical Engineering of the Universitat Rovira I Virgili,

CERTIFY:

That the present study, entitled "ENVIRONMENTAL RISK ASSESSMENT IN THE MEDITERRANEAN REGION USING ARTIFICIAL NEURAL NETWORKS" presented by Marelys Josefina Mujica Chacín, in partial fulfillment of the requirements for the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of this university and meets the requirements to obtain the European Mention.

Tarragona, February 20th 2012.

Dr. Francesc Giralt i Prat

Dr. Robert Rallo Moya

A Diego,

Ana Mercedes y Alba Cecilia

Acknowledgments

I would like to thank to my supervisors Dr. Francesc Giralt and Dr. Robert Rallo for their guidance, support and advices. Also, special thanks to Dr. Gabriela Espinosa for her support, guidance and help during the first years of this project.

To my PhD partners, my lab's colleagues and the personnel of chemical engineering department of URV for the nice talks during coffee breaks and "hall-talks". Been grateful by all good friends that URV doctoral program made me find and who will remain forever… mi familia de Tarragona!

Special thanks to my family, my parents Miguel and Etel, my brothers Migue and Fernando and my sister Gaby for encourage me to successfully finish this thesis.

Also, I'm infinitely grateful of my lovely husband Diego and my daughters Ana Mercedes and Alba Cecilia for their unconditional support and patience during my doctoral studies.

For the financial support, I want to acknowledge to the European Union (NoMiracle Project, European Commission, FP6 Contract No. 003956), and Servei de Gestió de la Recerca and Agència de Gestió d'Ajuts Universitarisi de Recerca (AGAUR) of Generalitat de Catalunya.

Resumen

Existe una creciente preocupación de nuestra sociedad sobre aspectos medioambientales como el cambio climático, la pérdida del hábitat, la deposición ácida, el descenso de la diversidad biológica y los impactos ecológicos de compuestos contaminantes como pesticidas y químicos tóxicos. El análisis de riesgo medioambiental puede ayudar a identificar estos problemas, establecer prioridades y proveer bases científicas para acciones regulatorias.

La realización de un estudio satisfactorio de estos y otros problemas medioambientales, requiere el uso de herramientas inteligentes que sean capaces de visualizar y extraer relaciones entre fuentes antropogénicas y sus efectos sobre el medio. Los desafíos principales relacionados a la aplicación de éstas técnicas son la naturaleza altamente no lineal de las relaciones causa – efecto estudiadas, así como la falta de información geográfica confiable. Los algoritmos de aprendizaje de máquinas y las técnicas de búsqueda de datos han probado ser muy exitosos en aplicaciones que incluye el manejo de datos de altas dimensiones y la presencia de incertidumbre.

Los mapas auto-organizados han demostrado ser una herramienta apropiada para la clasificación y visualización de grupos de datos complejos. Redes neuronales, como los mapas auto-organizados (SOM) o las redes difusas ARTMAP (FAM), se utilizan en este estudio para evaluar el impacto medio ambiental acumulativo en diferentes medios (aguas subterráneas, aire y salud humana). Los SOMs también se utilizan para generar mapas de concentraciones de contaminantes en aguas subterráneas simulando las técnicas geostadísticas de interpolación como kriging y cokriging. Para evaluar la confiabilidad de las metodologías desarrolladas en esta tesis, se utilizan procedimientos de referencia como puntos de comparación: la metodología DRASTIC para el estudio de vulnerabilidad en aguas subterráneas y el método de interpolación espacio-temporal conocido como Bayesian Maximum Entropy (BME) para el análisis de calidad del aire.

Esta tesis contribuye a demostrar las capacidades de las redes neuronales en el desarrollo de nuevas metodologías y modelos que explícitamente permiten evaluar las dimensiones temporales y espaciales de riesgos acumulativos, tanto para receptores humanos como ecológicos. Los mapas auto-organizados y las redes difusas proveen un marco consistente que pueden ser aplicados en diferentes niveles de la metodología del análisis de riesgo medioambiental aplicado en diferentes resoluciones espaciales.

Summary

Our society is increasingly aware of environmental issues including climate change, habitat loss, acid deposition, a decrease in biological diversity, and the ecological impacts of xenobiotic compounds such as pesticides and toxic chemicals. Environmental risk assessment (ERA) can help identify these environmental problems, establish priorities, and provide a scientific basis for regulatory actions.

A successful assessment of these and other environmental problems requires the use of intelligent tools capable of visualizing and extracting causal relationships among stressors and their effects. The main challenges related to the application of these techniques are the highly non-linear nature of the relationships between stressor and their effects as well as the lack of reliable and geographically distributed data. Machine learning algorithms and data-mining techniques have proven to be very successful in applications that include the management of high dimensional datasets and the presence of uncertainty.

The self-organizing map algorithm has demonstrated to be an appropriate tool for the classification and visualization of complex datasets. Neural networks, such as Self-Organizing Maps (SOM) or Fuzzy ARTMAP (FAM), are used to address cumulative environmental impact in different media (groundwater, air and human health). SOMs are also used to generate pollutant concentration maps in groundwater by simulating the kriging and co-kriging interpolation methodology. In order to evaluate the reliability of the methodologies developed in this thesis, well-known reference approaches were performed for comparison purposes: DRASTIC index for groundwater vulnerability assessment and Bayesian Maximum Entropy (BME) for air quality assessment.

This thesis contributes to demonstrate the capabilities of artificial neural networks in the development of new methods and models that explicitly address temporal and spatial dimensions of cumulative risks, both for human and ecological receptors. The Self-Organizing Map (SOM) algorithm and Fuzzy ARTMAP neural classifier provide a consistent framework that can be successfully applied at different levels of the ERA framework and at different spatial resolutions.

Contents

List of Figures

List of Tables

List of Abbreviations

%s	Land surface slopes
A	Aquifer media – DRASTIC rated
ACA	Catalan water agency (Agència Catalana de l'Aigua)
Al	Aluminium
ANN	Artificial Neural Network
ART	Adaptive Resonance Theory
As	Arsenic
Ba	Barium
BME	Bayesian Maximum Entropy
BMU	Best Matching Unit in a trained self-organizing map
C	Hydraulic conductivity of the aquifer – DRASTIC rated
Cd	Cadmium
CDF	Cumulative Distribution Function
CHEBRO	Ebro river agency (Confederación Hidrográfica del Ebro)
C-planes	Components planes of trained self-organizing map
Cr	Chrome
Cu	Cooper
D	Depth to water table – DRASTIC rated
EPA	United States Environmental Protection Agency
ERA	Environmental Risk Assessment
EU	European Union
FAM	Fuzzy ART Map neural network
Fe	Lead
FFN	Feed-Forward neural Network
GENCAT	Catalonia government (Generalitat de Catalunya)
GIS	Geographic Information Systems
H	Piezometric level
HHRA	Human Health Risk Assessment
I	Impact of vadose zone – DRASTIC rated
ICC	Catalan cartographic institute (Institut Cartogràfic de Catalunya)
IGME	Spanish geologic institute (Instituto Geológico y Minero de España)
IK	Indicator Kriging
Ka	Hydraulic conductivity of the aquifer
Ks	Soil's permeability
land	Soil uses map (land uses)
MAE	Mean arithmetic error
Max	Maximum value
Mean	Mean arithmetic value
Min	Minimum value
Mn	Manganese

Mo	Molybdenum
mse	Mean standard error
Ndm	Number of daily measurements
Ni	Nickel
NO_2	Nitrate
NO_3	Nitrite
OK	Ordinary Kriging
O_3	Ozone
P	Annual rainfall
Pb	Lead
PDF	Probability Density Function
PRTR	Spanish register of emissions and pollutant sources
PM	Particulate Matter
PM_{10}	Particulate matter of diameter bellow than 10 micrometers
$PM_{2.5}$	Particulate matter of diameter bellow than 2.5 micrometers
qe	Quantization error of SOM
Q1	25^{th} quartile
Q3	75^{th} quartile
R	Net recharge – DRASTIC rated
RBFN	Radial Basis Function neural Network
RNN	Recurrent Neural Network
S	Soil media – DRASTIC rated
ST	Spatio-temporal
S/TRF	Space/Time Random Field
Sb	Antimony
Se	Selenium
SK	Simple Kriging
SO_4	Sulfates
SOM	Self-Organizing Map
std	Standard deviation
T	Topography – DRASTIC rated
te	Topology error of the SOM
T-SOM	Temporal Self-Organizing Map
TSP	Total Suspended Particles
UK	Universal Kriging
UB	Universitat de Barcelona
U-matrix	Unified distance matrix of trained self-organizing map
US	United States
UTM	Universal Transverse Mercator coordinate system
vIndex	Vulnerability index
wIndex	Weighted vulnerability index
Zn	Zinc

Chapter 1

Introduction

Our society is increasingly aware of environmental issues including climate change, habitat loss, acid deposition, a decrease in biological diversity, and the ecological impacts of xenobiotic compounds such as pesticides and toxic chemicals. Environmental risk assessment can help identify these environmental problems, establish priorities, and provide a scientific basis for regulatory actions.

A successful assessment of these and other environmental problems requires the use of intelligent tools capable of visualizing and extracting causal relationships among stressors and their effects. The main challenges related to the application of these techniques are the highly non-linear nature of the relationships between stressor and their effects as well as the lack of reliable and geographically distributed data. Machine learning algorithms and data-mining techniques have proven to be very successful in applications that include the management of high dimensional datasets and the presence of uncertainty.

The self-organizing map algorithm has demonstrated to be an appropriate tool for the classification and visualization of complex datasets. Neural networks, such as Self-Organizing Maps (SOM) or Fuzzy ARTMAP, are used to address cumulative environmental impact in different media (groundwater, air and human health). SOMs are also used to generate pollutant concentration maps in groundwater by simulating the kriging and co-kriging interpolation methodology. In order to evaluate the reliability of the methodologies developed in this thesis, well-known reference approaches were performed for comparison purposes: DRASTIC index for groundwater vulnerability assessment and Bayesian Maximum Entropy (BME) for air quality assessment.

This thesis contributes to demonstrate the capabilities of artificial neural networks in the development of new methods and models that explicitly address temporal and spatial dimensions of cumulative risks, both for human and ecological receptors.

1.1 Motivation

This PhD thesis was developed under the frame of the European Project "Novel Methods for Integrated Risk Assessment of Cumulative Stressors in Europe, NOMIRACLE". European Commission FP6 contract number 003956.The integrated project consortium counted scientists within human toxicology and epidemiology, ecotoxicology, environmental engineering, toxicogenomics, physics, mathematical modeling, geographic informatics, and social and economic sciences. Thirty eight institutions from 17 countries participated in the project.

The main objectives of NoMiracle project were:

1. To develop new methods for assessing the cumulative risks from combined exposures to several stressors including mixtures of chemical and physical/biological agents.
2. To achieve more effective integration risk analysis of environmental and human health effects.
3. To improve our understanding of complex exposure situations and develop adequate tools for sound exposure assessment.
4. To develop a research framework for the description and interpretation of cumulative exposure and effect.
5. To quantify, characterize and reduce uncertainty in current risk assessment methodologies, e.g. by improvement of scientific basis for settings safety factors.
6. To develop assessment methods which take into account geographical, ecological, social and cultural differences in risk concepts and risk perceptions across Europe.
7. To improve the provisions for the application of the precautionary principle and to promote its operational integration with evidence based assessment methodologies.

The organization of the project was held on four research pillars:

Pillar 1: Risk scenarios

WP 1.1 Data background for scenario selection
WP 1.2 Scenario selection and ranking

Pillar 2: Exposure assessment

WP 2.1 Matrix-compound interaction
WP 2.2 Available exposure
WP 2.3 Metabolic fate

WP 2.4 Region specific environmental fate

Pillar 3: Effect assessment

WP 3.1 Interactive toxicology in diverse biological systems
WP 3.2 Combined effects of natural stressors and chemicals
WP 3.3 Toxico-kinetic modeling
WP 3.4 Molecular mechanisms of mixture toxicity

Pillar 4: Risk assessment

WP 4.1 New concepts for probabilistic risk assessment
WP 4.2 Explicit modeling of exposure and risk in space and time
WP 4.3 Dealing with multiple and complex risk
WP 4.4 Risk presentation and visualization

The present work was carried out in the context of work pillars 4.2 and 4.4 of the NoMiracle project. The aim to generate novel methods to explicit modeling environmental exposure and risk in space and time, and the previous knowledge of the capabilities of artificial neural networks (ANN) to approximate accurately non-linear input-output relationships, were the main motivation to explore the non-linear correlations capabilities of ANNs in environmental risk assessment.

Data-driven techniques, as artificial neural networks, require a good and reliable data set in order to perform properly. The area selected for the study was the Mediterranean region of Catalonia in north-eastern Spain due the availability of a complete and public data set of geologic, climatologic and environmental variables. Main data sources were provided by Catalan and Spanish environmental agencies.

During the last years, several environmental-status studies over Catalonia region have been finished and there is a great concern on anthropogenic pollution in different media, like groundwater and air. As a matter of fact, Catalan environmental agency had been specially focused on nitrate pollution in groundwater bodies over Catalonia and some hydrogeological areas were labeled as protected water bodies due high levels of nitrate vulnerability. Based on public available data from Catalan government, some objectives and planning were developed for the development of this thesis as presented in next section.

1.2 Hypothesis and Objectives

In the framework of NoMiracle project, the need to develop new methods and models that explicitly address the temporal and spatial dimensions of cumulative risks, both for human and ecological receptors, the main hypothesis of this study stated that:

"A*rtificial neural networks are capable to generate cause-effect relationships between pollutants sources and/or concentrations in a media and human or ecological receptors throughout environmental risk assessment context".*

In order to verify the hypothesis of this project, the Mediterranean region of Catalonia was selected as study area and a general objective was formulated:

- Study of artificial networks capabilities to assess the environmental risk of cumulative pollution in the Mediterranean region.

Four scenarios were established to evaluate the capabilities of some artificial neural networks to assess environmental risk in the area of study: groundwater vulnerability assessment, lead exposure assessment, air quality assessment and human health risk assessment. Three artificial neural networks were selected to accomplish the general objective: Self-organizing map (SOM), Fuzzy ARTMAP neural network (FAM) and Backpropagation neural network (BP).

Specific objectives were stated in the basis of environmental risk scenarios selected for the study:

Objective 1: Study the capabilities of self-organizing maps to perform groundwater vulnerability assessment and generate reliable groundwater vulnerability maps using available hydrogeological and climate data in the area of study.

Objective 2: Study the capabilities of fuzzy ARTMAP and backpropagation neural networks to assess cause-effect relationships in the lead exposure assessment over the area of study.

Objective 3: Explore spatio-temporal interpolation capabilities of self-organizing maps to perform air quality assessment in the area of study.

Objective 4: Explore the capabilities of self-organizing maps to perform cause-effect relationships between air pollution and human health by respiratory diseases in the area of study.

1.3 Scientific contributions

Three peer-reviewed deliverables were generated during this study under the NoMiracle European project:

- CD: Fuzzy ARTMAP neural classifier for cluster analysis and demonstration of a variable selection approach, April 2006.
- A model for exposure and risk relationships using neural networks at different sites scales, May 2007.
- Report on spatio-temporal models in a Catalan county region, Jun 2008.

Several contributions to conferences or workshops were presented throughout the development of this thesis:

- Mujica M. Espinosa G; Guardiola X; Rallo R; Saiz M; Ferré J; Giralt F. Self-Organization of geostatistical information for vulnerability analysis SETAC Europe 16th Annual Meeting, The Hague, May 2006.
- Mujica M. Espinosa G; Grifoll J; Giralt F. Self-Organizing groundwater vulnerability maps. 5th European Congress on Regional Geoscientific Cartography and Information Systems (Econgeo2006), Barcelona, June 2006.
- Rallo, R., Mujica, M., Climent, J., Espinosa, G., Grifoll, J., Giralt, F. (Ecological) Risk Mapping based on self-organizing maps. 1st Open International NoMiracle Workshop. Ecological and Human Health Risk Assessment. Ispra, Italy , June 2006.
- Rallo R; Mujica M; Climent J; Espinosa G; Giralt F. Self-Organizing Maps and Gaussian Mixture Models for Environmental Risk Assessment and Mapping. SETAC Europe 17th Annual Meeting, Porto, May 2007.
- Mujica, M; Rallo, R; Giralt, F. SOM-based approach to assess the cumulative exposure and cardio-respiratory effects of particulate matter in Catalunya. NoMiracle Workshop on Cumulative Risk Assessment: a Challenge for Science and Management. Ravenstein, The Netherlands, March 2009.

Two peer-reviewed publications were generated during this thesis:

- Pistocchi A; Groenwold J; Lahr J; Loos M; Mujica M; Ragas A; Rallo R; Sala S; Schlink U; Strebel K; Vighi M; Vizcaino P. (2011) “Mapping Cumulative Environmental Risks: Examples from the EU NoMiracle Project”. Environmental Modeling & Assessment, 16: 119-133. (Annex A.1).
- Mujica, M.; Rallo, R.; Giralt, F. (2012) “Intrinsic Groundwater Vulnerability Assessment using Self-organizing Maps”. Environment International (to be submitted). (Annex A.2).

1.4 Organization of the manuscript

This document is organized in seven main chapters.

At the first chapter a general introduction and motivation for the thesis is presented. Hypothesis and objectives are stated and scientific contributions derived from this work are presented.

Second chapter presents the basic concepts of different techniques and tools used in the study. Concepts of vulnerability, exposure and risk assessment are stated in order to define the conceptual framework of the work. Reference methodology like DRASTIC vulnerability index for groundwater vulnerability is described in this chapter. Also, descriptions of well-known spatial and spatio-temporal interpolation techniques are presented. General descriptions of artificial neural networks employed in this work are also presented in this chapter.

Third chapter address the vulnerability assessment applied for groundwater resources in Catalonia region as stated in objective 1. Self-organizing map neural classifier was selected to study the capabilities of this neural network to perform reliable groundwater vulnerability maps.

Exposure to lead assessment for Catalonia region is presented at chapter four. In this chapter, objective 2 is addressed by the evaluation of cause-effect capabilities of fuzzy ARTMAP and backpropagation neural networks.

At chapter five, spatio-temporal air quality assessment for Catalonia region is presented using self-organizing maps and bayesian maximum entropy spatio-temporal interpolation techniques (objective 3).

Chapter six presents the exploration of self-organizing maps capabilities to generate cause-effect relationships between air pollution in Catalonia and human health effects like respiratory diseases, in order to accomplish objective 4 of this project.

Finally, at chapter seven, the general conclusions of the developed work are presented Also, some future opportunities of data-driven research opportunities using artificial neural networks that have been detected during the development of this thesis are presented.

Chapter 2

Background Concepts

Concepts for vulnerability, exposure and risk are stated in the following definitions in agreement with the conclusions and recommendations of the International Conference on Vulnerability of Soil and Groundwater to Pollutants (Duijvenbooden and Waegeningh 1987) and the work of Lahr and Kooistra (2010):

Vulnerability: "Use the presence and geographical distribution of sensitive receptors of stress to map more and less vulnerable areas". In the context of groundwater vulnerability, two definitions have to be stated:

Intrinsic vulnerability: "Assessment of vulnerability areas based on intrinsic hydrogeological characteristics of aquifers and overlying media".

Specific vulnerability: "Assessment of vulnerability areas based on intrinsic hydrogeological characteristics of aquifers, overlying media and specific characteristics of the pollutant or pollutants over the area of study".

Exposure: "Combine measured or predicted environmental contamination levels with the geographical distribution of an (ecological or human) exposure receptor".

Risk (General): "Compare (measured or predicted) environmental concentrations to simple environmental threshold levels (environmental quality criteria). Map results of extensive modeling/simulation of contamination, exposure and effects (model train approach). Combine maps of vulnerability and maps of (potential) impact/environmental pressures".

Risk of multiple stressors: "Calculate combined risk from individual risk parameters for single stressors using certain principles and algorithms. Use multivariable statistical analysis to reduce dimensionality and map statistical parameter values."

2.1 Environmental risk assessment

Environmental risk assessment (ERA) can be defined as a procedure that defines the probability and magnitude of adverse effects to human health and/or natural resources posed by environmental agents. ERA covers the risk to ecosystems (ecological risk assessment) like air, water, land and biological species, and risk to humans, exposed or impacted (human health risk assessment) as defined by European Environment Agency (EEA, 1998).

Principal steps in an ERA can be summarized as following:

- Hazard identification: this is the problem formulation step; includes identification of the property or situation that could lead to harm.
- Consequences identification: this is the identification of possible consequences if the hazard occurs.
- Magnitude of consequences: this step includes the spatial and temporal scale of the consequences and the time required to onset the consequences.
- Probability of consequences: this means the estimation of the probability of occurrence of the consequence; the probability of the receptors being exposed and the probability of harm as a consequence of exposure to hazard.
- Risk estimation: evaluation the significance of a risk.

ERA has gained an important place in the evaluation of the environmental impact of different stressors to human and/or ecological receptors. Jones (2001) for example study the impact of climate change on different units detected as vulnerable using ERA framework. Many applications of ERA on pharmaceutical drugs substances impact have been published (Carlsson et al., 2006; Escher et al., 2011; Ginebreda et al., 2010; Santos et al., 2007). Industrial and engineered compounds and materials have also been evaluated by ERA to human health effects or ecological adverse effects (Escher and Fenner, 2011; Savolainen et al., 2010).

The ERA methodology gives a general framework to perform specific risk assessments over ecological and/or human receptors that would help governments and decision-makers in the management of policies and strategies to preserve and/or improve environmental status of an area (Lahr and Kooistra, 2010). In this work, ERA was used as reference framework for the development of methodologies and models at the study area selected.

2.2 Groundwater vulnerability

The assessment of groundwater vulnerability is usually performed on the basis of vulnerability indicators reflecting individual factors affecting vulnerability, combined in order to obtain a comprehensive and synthetic characterization of the actual aquifer vulnerability. Groundwater vulnerability studies have been developed all-world around in order to identify susceptible zones to pollution that should be protected or remediated. Different techniques have been implemented to generate vulnerability maps, the most popular are the overlay and index methods, followed by statistical methods and others.

- *Overlay and Index methods:* Are simple mathematical models, consisting in algebraic operations of hydrogeological parameters. Table 2.1 presents a summary of most used vulnerability index found in literature.Advantages: Are algebraic models, and are easy to implement in GeographicalInformationSystemssoftware (GIS). Disadvantages: Are simple mathematical representations of expert opinion and not on process representation. Frequently there are some difficulties to find the specific data required for each model.
- *Fuzzy methods:* Estimateweights and rates for Index methods using fuzzy rules. (Dixon, 2005a,b; Gemitzi et al., 2006; Mao et al., 2006; Mazari Hiriart et al., 2006; Uricchio et al., 2004).Advantages: Reduces uncertainty in weighting and rating process. Disadvantages: Requires expertise in fuzzy logic algorithms and expert criteria to select the appropriate fuzzy rules.
- *Statistical methods:* Estimate groundwater vulnerability by statistical analysis of point pollution data. Quantify vulnerability by determining the statistical dependence between observed contamination, observed environmental conditions and observed land uses that are potential sources of contamination. (Panagopoulos et al., 2006; Worrall and Besien, 2005; Worrall and Kolpin, 2003).Advantages: Consider pollution data. Provide measure of uncertainty in the model. Disadvantages: Site specific and difficult to develop.
- *Process-based simulation models:* Use the governing equations for water flow and solute transport. The focus is on computing travel times or concentrations of a contaminant in the unsaturated and groundwater zones. (Lindström, 2005).Advantages: Quantitative point of view. Models can be used to study process in generic hydrogeological settings. Disadvantages: Time consuming. Extensive data input requirements and require advanced expertise in the process involved.

Table 2.1. Overlay and index methods for groundwater vulnerability assessment

METHOD	CHARACTERISTICS	ADVANTAGES	DISADVANTAGES	REFERENCES
DRASTIC	7 parameters: Depth to water table, Net recharge, Aquifer media, Soil media, Topography, Impact of vadose zone, Hydraulic conductivity.	Fast assessment.	Rating and weights (expert criteria). Correlation between parameters.	(Aller et al., 1987)
SINTACS	7 parameters: Depth to water table, Net recharge, Aquifer media, Soil media, Topographic surface, Unsaturated zone, Hydraulic conductivity.	Fast assessment.	Rating and weights (expert criteria). Specific software required.	(Civita, 1994)
SEEPAGE	6 parameters: Soil slope, Depth to water, Vadose zone material, Aquifer material, Soil depth and Attenuation potential (texture of surface soil, texture of subsoil, surface layer pH, organic matter content of the surface, soil drainage class, soil permeability).	Fast assessment.	Rating and weights (expert criteria). Requires detailed geological information.	(Moore and John, 1990)
EPIK	4 parameters: Epikarst, Protective cover, Infiltration conditions, Karst network development.	Specific for karstic aquifers. Fast assessment.	Rating and weights (expert criteria).	(Doerfliger and Zwahlen, 1997)
PI	2 parameters: Protective cover and Infiltration conditions.	Fast assessment.	Expert criteria.	(Goldscheider, 2005)
COP	3 parameters: Flow concentration, Overlying layers and Precipitation.	Specific for karst aquifers. Uses different level of available data.	Only Rating (expert criteria). Complex algorithm to score parameters.	(Vías et al., 2006)
GOD	3 parameters: Groundwater occurrence, Overall aquifer class, Depth to water table	Fast assessment.	Rating and weights (expert criteria). Does not consider soil media properties.	(Foster, 1987)
AVI	2 parameters: Thickness of each sedimentary layer above the uppermost saturated aquifer and Hydraulic conductivity of each sedimentary layer.	Does not consider ratings and/or weights.	So few hydrogeological parameters.	(Van Stempoort et al., 1993)
German method (Hölting method)	5 parameters: Soil, Deeper subsoil, Seepage water, Lump sum addition for perched aquifer situation and Lump sum addition for artesian groundwater condition.	Fast assessment.	Parametric method not weighting. Requires detailed geological information.	(Burchart et al., 2006)

2.2.1 DRASTIC vulnerability index

The DRASTIC Index was developed by the EPA (Environmental Protection Agency of United States of America) to be a standardized system for evaluating groundwater vulnerability to pollution and has been used worldwide (Aller et al., 1987). The method has four main assumptions:

- the pollutant is introduced at the ground surface
- the contaminant flushes to the groundwater by precipitation
- the contaminant has the mobility of water
- the study area should be 100 acres or larger

DRASTIC Index considers seven hydro-geological properties, as following:

Depth to water table (D): is the distance from the surface to the water table. It is evident that the shallower the depth, the more vulnerable is the aquifer to pollution.

Net recharge (R): is the total quantity of water per unit area which reaches the water table. Recharge is the main vehicle for leach and pollutant transport to the water table. The vulnerability to contamination is enhanced by highs recharge rates.

Aquifer media (A): refers to the properties of the rock that serves as an aquifer. Lithology and grain size is determinant for the transport of pollutants within the aquifer. The property of a rock to be pervaded by a fluid is called permeability. Also porosity plays an important role in vulnerability assessment. The higher the permeability and porosity in the aquifer, the higher the vulnerability to contamination.

Soil media (S): is referred to the upper weathered zone of the earth, the first 1.5 meters from the ground surface. The content of clay and organic matter are relevant in controlling the pollutant infiltration to aquifers. In general, presence of clay and small grain size reduces the vulnerability of the groundwater to pollution.

Topography (T): refers to the slope of the land surface. Vulnerability to contamination is reduced as the slopes increases due the increment in the runoff capacity of the media.

Impact of the vadose zone (I): refers to the unsaturated zone above the water table. Like the soil media, the texture of the vadose zone determines the travel time of pollutant through.

Hydraulic conductivity of the aquifer (C): refers to the rate at which water flows horizontally through an aquifer. Vulnerability is increased as hydraulic conductivity increases.

The DRASTIC vulnerability mapping involves the overlaying the seven hydrogeological properties as described in equation 2.1.

$$\text{DRASTIC Index} = DrDw + RrRw + ArAw + SrSw + TrTw + IrIw + CrCw \quad (2.1)$$

r: rating, w:weight

Each DRASTIC feature is assigned a weight relative to each other in an increasing range of importance from 1 to 5 based on expert criteria. Table 2.2 shows DRASTIC weights formulated by Aller et al. (1987) for a general aquifer (intrinsic vulnerability) and for an aquifer exposed to pesticide pollution (specific vulnerability). There are many others contribution to generating specific site weights taking into account land use or agricultural activities (Secunda et al., 1998; Umar et al., 2009).

Ratings for each DRASTIC variable are assigned a value between 1 and 10, in an increasing order of impact to vulnerability. This step requires a laborious pre-processing task for each variable, based on expert criteria and knowledge of the process under study.Piscopo (2001) presented a general methodology to generate DRASTIC features from common hydrogeological data that is used in this work and presented in section 3.3.2 of chapter 3 of this manuscript.

Table 2.2. DRASTIC weights by Aller et al. (1987)

Feature	General	Pesticide
Depth to water	5	5
Net Recharge	4	4
Aquifer media	3	3
Soil media	2	5
Topography	1	3
Impact of vadose zone media	5	4
Hydraulic Conductivity of aquifer	3	2

Many authors have applied DRASTIC methodology in different world-wide areas. For example, Rhaman (2008) applied DRASTIC index in a shallow aquifer in India, Senar et al. (2009) generated a DRASTIC based vulnerability map of the Senirkent-Uluborlu basin in Turkey and Bojórquez-Tapia et al. (2009) defined a visualization-DRASTIC for two urban watersheds in Mexico. The DRASTIC methodology was used for comparison purpose of the new vulnerability methodology developed in the study.

2.3 Spatial interpolations

2.3.1 Geostatistics

Geostatistics is a branch of applied statistics developed for the mining activities by Matheron (1963) as an application of the theory of random functions for estimating natural phenomena (Matheron, 1971; 1973; Samper and Carrera, 1996). Geostatistics was initially developed to be used in geo-sciences. Nowadays, geostatistics is used in petroleum geology, hydrogeology, meteorology, oceanography, geochemistry, forestry, environmental control, landscape ecology, agriculture, etc.

The basic concept of geostatistics is that of scales of spatial variation (Matheron, 1963). Spatially independent data show the same variability regardless of the location of data points. However, spatial data in most cases is not spatially independent. Geostatistics is based on the assumption that measurements lying closer together tend to be more alike than those farther apart. The exact nature of this pattern varies from data set to data set; each set of data has its own unique function of variability and distance between data points. This fundamental geographic principal is called spatial autocorrelation and can be examined by means of variogram analysis (Daly and Warren, 1998; Dixon, 2004).

Spatial variability is generally computed as a function called semivariance. Spatial autocorrelation can be analyzed using correlograms, covariance functions and variograms (initially called as semivariograms). Semivariance is a measure of the degree of spatial dependence between samples. The magnitude of the semivariance between points depends on the distance between the points. A smaller distance yields a smaller semivariance and a larger distance results in a larger semivariance. The plot of the semivariances as a function of distance from a point is referred to as a semivariogram. The semivariance increases as the distance increases until at a certain distance away from a point the semivariance will equal the variance around the average value, and will therefore no longer increase, causing a flat region to occur on the semivariogram called a sill. From the point of interest to the distance where the flat region begins is termed the range or span of the regionalized variable (Figure 2.1). Within this range, locations are related to each other, and all known samples contained in this region, also referred to as the neighborhood, and must be considered when estimating the unknown point of interest.

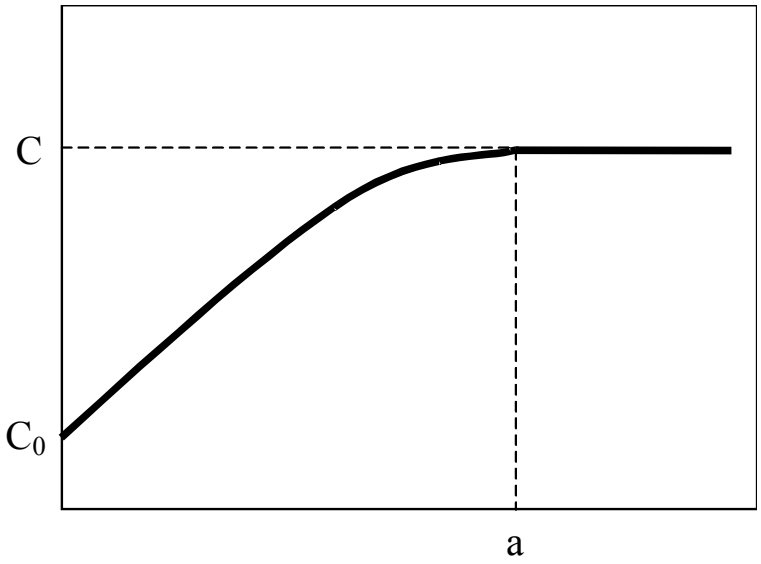


Figure 2.1. Theoretical Variogram parameters. Range (a), Sill (C) and Nugget (C_0)

The characteristic parameters of a variogram are:

Sill (C): The semivariance value at which the variogram levels off. Also used to refer to the "amplitude" of a certain component of the semivariogram.

Range (a): The lag distance at which the semivariogram (or semivariogram component) reaches the sill value. Presumably, autocorrelation is essentially zero beyond the range.

Nugget (C_0): In theory the semivariogram value at the origin (0 lag) should be zero. If it is significantly different from zero for lags very close to zero, then this semivariogram value is referred to as the nugget. The nugget represents variability at distances smaller than the typical sample spacing, including measurement error.

The semivariance function is represented by the following equation:

$$\gamma(h) = \frac{1}{2N_P(h)} \sum_{i=1}^{N_P(h)} \left(Z(x_i) - Z(X_i + h)\right)^2 \qquad (2.2)$$

$N_P(h)$: total pair numbers at distance h;
h: distance between pairs
Z(xi): experimental values at each location xi
xi: spatial locations

Sample variogram have to be fitted by a theoretical variogram in order to compute kriging interpolations. There are different models and can be used alone or as a combination of them. The most frequently used are the spherical, exponential and gaussian as shown in Figure 2.2 and presented in equations 2.3 to 2.5.

- *Spherical model*:

$$\gamma(h) = \begin{cases} C\left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & h \le a \\ C & h > a \end{cases} \qquad (2.3)$$

- *Exponential model*:

$$\gamma(h) = C\left[1 - \exp\left(-\frac{h}{a}\right)\right] \qquad (2.4)$$

$$|h| > 0$$

- *Gaussian model*:

$$\gamma(h) = C\left[1 - \exp\left(-\frac{h^2}{a^2}\right)\right] \qquad (2.5)$$

$$|h| > 0$$

The spherical model actually reaches the specified sill value, C, at the specified range, *a*. The exponential and Gaussian approach the sill asymptotically, with *a* representing the practical range, the distance at which the semivariance reaches 95% of the sill value. The Gaussian model, with its parabolic behavior at the origin, represents very smoothly varying properties. (However, using the Gaussian model alone without a nugget effect can lead to numerical instabilities in the kriging process). The spherical and exponential models exhibit linear behavior the origin, appropriate for representing properties with a higher level of short-range variability.

Kriging is the estimation procedure used in geostatistics using known values and a semivariogram to determine unknown values. It was named after D. G. Krige from South Africa (Krige, 1951). The procedures involved in kriging incorporate measures of error and uncertainty when determining estimations. Based on the semivariogram used, optimal weights are assigned to unknown values in order to calculate unknown ones. Since the variogram changes with distance, the weights depend on the known sample distribution. Different kriging approaches based have been developed on the definitions of variance estimator. Most popular kriging models are presented in the following lines.

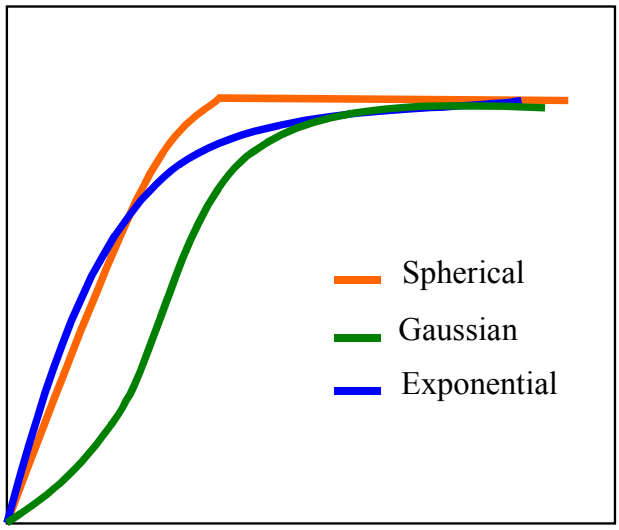


Figure 2.2. Theoretical variogram models (Spherical, Gaussian and Exponential)

Ordinary Kriging (OK) is a variety of kriging which assumes that local means are not necessarily closely related to the population mean, and which therefore uses only the samples in the local neighborhood for the estimate. Equations 2.6 to 2.8 present OK mathematical definition.

$$Z^*(v) = \sum \lambda_i Z(x_i) \tag{2.6}$$

$$\sum \lambda_i \gamma(x_i, x_j) + \mu = \gamma(x_j, \mu) \tag{2.7}$$

$$\sigma^2 = \sum \lambda_i \gamma(x_i, v) - \gamma(v, v) + \mu \tag{2.8}$$

v: spatial location for prediction
$Z^*(v)$: estimated value by kriging
μ: sample mean
σ^2: estimation variance
λ_i: kriging weights

Simple kriging(SK) uses the average of the entire data set, while ordinary kriging uses a local average (the average of the scatter points in the kriging subset for a particular interpolation point). As a result, simple kriging can be less accurate than ordinary kriging, but it generally produces a result that is "smoother" and more esthetically pleasing.

*Block Kriging*estimates the value of a block from a set of nearby sample values using kriging.

*Point Kriging*estimates the value of a point from a set of nearby sample values using kriging. The kriged estimate for a point will usually be quite similar to the kriged estimate for a relatively small block centered on the point, but the computed kriging standard deviation will be higher. When a kriged point happens to coincide with a sample location, the kriged estimate will equal the sample value.

Universal Kriging (UK) is a procedure similar to that of punctual kriging, but used when a trend, or slow change in average values, in the samples exists.

Indicator kriging (IK) is a geostatistical approach to geospatial modeling. Like OK, the correlation between data points determines model values. However, IK makes no assumption of normality and is essentially a non-parametric counterpart to OK. Instead of assuming a normal distribution at each estimate location, IK builds the cumulative distribution function (CDF) at each point based on the behavior and correlation structure of indicator transformed data points in the neighborhood (Lloyd and Atkinson, 2001). To achieve this, IK needs a series of threshold values between the smallest and largest data values in the set. These threshold values, referred to here as *IK cutoffs,* are used to numerically build the CDF of the estimation point. For each IK cutoff, data in the neighborhood are transformed into 0s and 1s: 0s if the data are greater than the threshold, and 1s if they are less. IK then estimates the probability that the estimation point is less than the threshold value, given this neighborhood of transformed data and a model of the IK cutoff correlation structure. Performing this operation for each cutoff across the range of data approximates the CDF at the estimation point. After the CDF is built, it must be post processed to produce probability maps for estimation maps and risk maps.

2.3.2 Bayesian maximum entropy

The Bayesian Maximum Entropy (BME) method has been successfully used in the mapping analysis of environmental contaminants in groundwater, surface water and ambient air (Vyas and Christakos, 1997; Christakos and Serre, 2000; Serre and Christakos, 2004). BME is based on modern spatio-temporal geostatistics (Christakos, 2000), so that classical geostatistics is a specific case of the former. There are several coordinate systems that can be used in BME compared with the Euclidean system on which 'classical' geostatistics is based. In this chapter a brief review of BME general principle is presented.

2.3.2.1 Spatio-temporal random field representation

In these studies a space/time random field (S/TRF) Z(***p***) is used to represent the uncertainties and natural variability associated with a contaminant Z at some space/time point ***p***(s,t), where *s* is the geographical coordinate and*t* is the time.

The total physical knowledge (*K)* available regarding a pollutant distribution is assumed to be composed by two main knowledge sources covering both, the general (G) and case-specific (S) knowledge, so that:

$$K = G \cup S \tag{2.9}$$

The G refers to background knowledge that includes physical laws, structured patterns and assumptions, and statistical moments. Some examples of G base are given in Table 2.3.

Table 2.3. Examples of general knowledge G bases

	G_α
Covariance	$\overline{\left(x_i - \overline{x_i}\right)\left(x_j - \overline{x_j}\right)}$
Variogram	$\frac{1}{2}\overline{\left(x_i - x_j\right)^2}$
λ-th order moment	$(x_i - \overline{x_i})^\lambda$ λ=1, …, L

The S refers to specific situation of the study case, includes empirical observations, site-specific data, etc. As a matter of fact, forPM_{10} mapping purposes, the specific knowledge base, S, will denote physical data ,x_{data}, obtained at points $\mathbf{p}_i$(i=1, …, m) for specific X, Y location coordinates. The S base is formed by two main contributions:

$$S: x_{data} = (x_{hard}, x_{soft}) \tag{2.10}$$

Hard data are exact measurements of pollutant concentrations. In contrast, *soft data* may include uncertain evidence about the random vector, x_{soft}. The x_{soft}could be expressed in terms of intervals or in a probabilistic form (Christakos and Serre, 2000).

2.3.2.2 Bayesian maximum entropy estimates

The integration and processing of the physical knowledge based on *K* leads to the posterior BME probabilistic density function (PDF*)* $f_k(x_k)$which, in turn, provides the full stochastic description of pollutant concentration (PM_{10}) at any estimation point $\mathbf{p}_k$ of interest. The BME posterior PDF $f_k(x_k)$ can be expressed as (Christakos and Serre, 2000), equation 2.11,

$$f_k(x_k) = A^{-1}\int dx_{soft} f_S(x_{soft}) f_G(x_{hard}, x_{soft}, x_k) \tag{2.11}$$

Where x_{hard},$x_{soft,}$ $x_k$represent the S/TRF for the pollutant at the hard, soft and estimation points, respectively, $f_S(x_{soft})$ is a PDFdescribing the uncertainty for the pollutant at the soft data points, $f_G(x_{hard}, x_{soft,}x_k)$ is a PDF integrating general knowledge about the pollutant and A is a normalization constant. From the posterior PDF$f_k(x_k)$ we can obtain several estimators of the pollutant concentration at the estimation point $\mathbf{p}_k$. BME allows the flexibility to choose among several estimators, and the choice of the best suited estimate will depend on the individual characteristics of each mapping situation. A common estimator used in environmental mapping analysis is the BME median estimator, $\overline{x}_{k,median}$, defined as the 50% quantile of the BME posterior PDF, and given by the following equation:

$$\int_{-\infty}^{\hat{x}_{k,mean}} dx_k x_k f_k(x_k) = 0.5 \tag{2.12}$$

Furthermore, because of the inherent randomness of the pollutant transport processes in air and the physical data inaccuracies, it is essential to assess the uncertainty associated with the estimated values (Christakos and Serre, 2000; Christakos et al., 2002)

The variance of the BME posterior PDF provides a useful assessment of the estimation accuracy for the BME median estimator. The BME posterior variance is denoted by $\sigma^2_{k/k}$ and is given by:

$$\sigma^2_{k/k} = dx_k (x_k - \overline{x}_{k,mean})^2 f_k(x_k) \tag{2.13}$$

This quality measure corresponds to the variance of the estimation error. However, unlike the classical kriging variance that is independent of the data values (the kriging variance is only covariance and spatial data configuration dependent, must not be confused with the variance of the kriging predictor), the $\sigma^2_{k/k}$ depends on the specific set of data values considered. In some cases, a more realistic assessment of the mapping error when the posterior PDFhas a complicated shape may be achieved using the concept of BME confidence sets, which are an extension of the classical confidence interval.

2.4 Artificial neural networks

Artificial Neural Networks (ANN) are mathematical models inspired by the structure and functionality of biological neural networks (Hagan, et. al., 1996). A neural network consists of a number of simple processing units called neurons. Each neuron is connected to others by a weightedlink (w_{ij} in Figure 2.3).

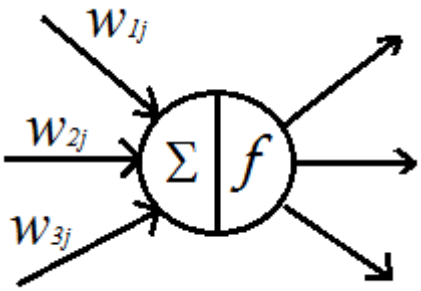


Figure 2.3. Simple neuron model in an ANN

ANN can be thought of as a model which approximates a function of multiple continuous inputs and outputs. The network consists of a *topology* graph of neurons (Figure 2.3), each of which computes a function (called an activation or transfer function, *f*) of the inputs carried on the in-edges and sends the output on its out-edges. Principal activation functions are summarized in Table 2.4. The inputs and outputs are weighed by weights (w_{ij}) and shifted by *bias* factor (b_i) specific to each neuron. The *bias* is an activation threshold that allows shifting of the activation function to the left or to the right.

The output of a neuron is computed by the following equation:

$$O_j = f\left(\sum_{i=1}^{n} w_{ij}x_i + b_j\right) \tag{2.14}$$

O_i is the output of the *j*th neuron, *f* is the activation or transfer function of the neuron, b_j is the bias of *j*th neuron, w_{ij} is the synaptic weight corresponding to the *i*th synapse of *j*th neuron, x_i is the *i*th input signal to *j*th neuron and n is the number of input signals to the *j*th neuron.

As their analogy to human brain, ANN requires an algorithm of training (learning process) which when given a function *f*, learns a set of weights and biases, which accurately approximate the function.

Table 2.4. Activations functions used in ANN

Name	Formula
Identity	$f(x) = x$
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Tanh	$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
Step	$f(x) = \begin{cases} -1 \; if \; x < 0 \\ 1 \; if \; x \geq 0 \end{cases}$

Different algorithms of "learning" defines different types of ANN, the principals are:

- Feed-forward neural networks (FFN)
- Radial basis function networks (RBFN)
- Self-organizing maps (SOM)
- Recurrent neural networks (RNN)
- Fuzzy ARTMAP Networks

2.4.1 Feed-forward neural networks

A feedforward neural network is an artificial neural network where the information moves only in one direction, from input to output nodes without any cycle or loop in the network.

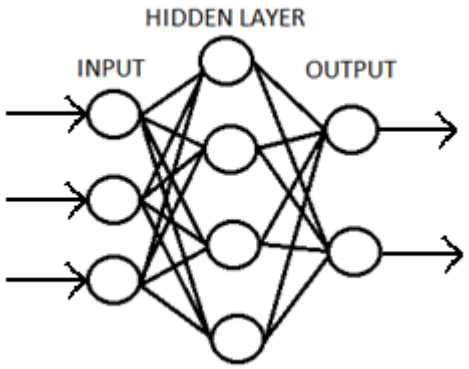


Figure 2.4.Multilayer perceptron feed-forward neural network architecture

A multilayer perceptron is a feedforward ANN that consists of multiple layers of nodes (neurons) connected to each other, as seen in Figure 2.4. The training algorithm used by this ANN is called backpropagation (backward propagation of errors) (Parker, 1985; Rumelhart et al., 1986). Backpropagation networks (BP) adjust their internal parameters with basis on the comparison between target and predicted values. At every epoch (training time step), the input data set is presented to the network (propagation step) and the predictions are compared to their corresponding target values; the inner parameters

of the network (weights and biases) are updated (updating step). In the training phase the error in the output node *j* in the *n*th data point is represented by

$$e_j = d_j(n) - y_j(n) \tag{2.15}$$

Where *d* is the target value and *y* is the value produced by the neuron. Corrections to the weights of the neurons are made based on minimization of the error in the entire output layer (equation 2.16). Minimization by gradient descent gives the change at each weight (equation 2.17) where y_i is the output of the previous neuron and α is the learning rate, which is selected to ensure fast and non-oscillatory convergence of the weights (typically ranges between 0.2 to 0.8). The errors propagate backwards from the output nodes to the inner nodes, and this can be accomplished by two methods: on-line learning and batch learning. In on-line learning, the updates of weightsare performed at each propagation stepand requiresmore updates than the batch algorithm. In batch learning the updating of weights is performed after a certain number of propagation stepsand requires more memory capacity than the previous one. Some limitations of backpropagation learning algorithm are that convergence is not guaranteed and (if existing) is very slow, also local minima can be found and normalization of input is required.

$$\varepsilon(n) = \frac{1}{2}\sum_j e_j^2(n) \tag{2.16}$$

$$\Delta w_{ij}(n) = -\alpha \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) \tag{2.17}$$

2.4.2 Radial basisfunction networks

A radial basis function network (RBFN) is a two-layer ANN that uses radial basis functions as activation functions in its hidden layer, and linear functions in its output layer (Lo, 1998). Types of radial basis functions are summarized in Table 2.5. Gaussian bell functions are the most used activation function in RBFN. Most applications of RBFN are in the field of function approximation, time series prediction and process control.

Table 2.5. Radial basis functions

Name	Formula
Gaussian	$f(x) = e^{-(\varepsilon r)^2}$
Multiquadric	$f(x) = \sqrt{1 + (\varepsilon r)^2}$
Inverse quadratic	$f(x) = \frac{1}{1 + (\varepsilon r)^2}$
Inverse Multiquadric	$f(x) = \frac{1}{\sqrt{1 + (\varepsilon r)^2}}$

2.4.3 Self-organizing maps

The self-organizing map (SOM)(Kohonen, 1990) is an unsupervised neural network algorithm that visualizes, clusters and classifies high dimensional data. A SOM consists of a number of neurons placed in a 2 or 3 dimensional grid. The neurons are connected to each other by a neighborhood relationship that governs the structure of the map. The grid can be hexagonal or rectangular, as shown in Figure 2.5. The shape of the grid is also an input parameter of the SOM, and a periodic shape such as torus (toroidal SOM) is commonly used to avoid border effects. (Rallo, 2007).

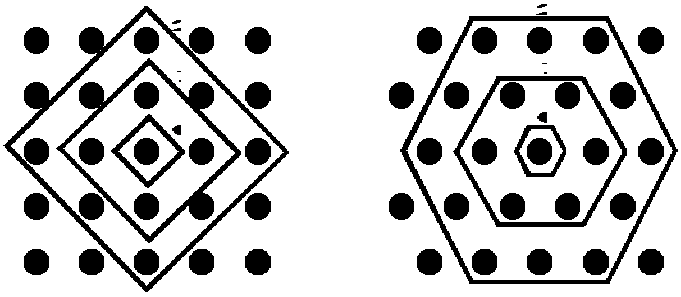


Figure 2.5.Neighborhoods (levels 0, 1 and 2) of unit (0) in (left) rectangular and (right) hexagonal lattices

The SOM algorithm is based on competitive learning (Kangas et al., 1990; Kaski, 1997; Kohonen, 1990) where neurons gradually become sensitive to different input categories in a domain of the input space. Each neuron is represented by a prototype vector $m_i=[m_{i1}, m_{i2}, ..., m_{in}]$ where n is the dimension of the input space. In the sequential algorithm the process begins with a random selection of one data point P and computing distances between P and every node of the SOM; the closest node to P (named the winner unit or Best Matching Unit, BMU) is moved the most in the direction of P, while the adjacent nodes are moved depending on their distance from the winner unit in the initial geometry. This preserves the topology or structure of the map, as illustrated in Figure 2.6.

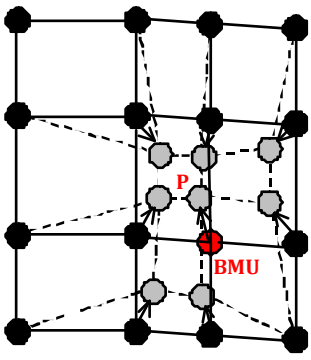


Figure 2.6. Best matching unit (BMU) and its topological neighbors (black dots) are updated during the SOM training process. Black and gray circles depict changes in location caused by the updating process

The update procedure displaces the BMU and neighbors towards the sample vector *P*. The weights vectors define the neurons or map units located in the crossing of the solid lines in Figure 2.6. This updating process, shown as dashed lines, preserves the topology or structure of the map. The update rule for the prototype vector of unit *i* is given by,

$$m_i = m_i + \alpha(t) h_{ci}(t)(P - m_i) \qquad (2.18)$$

where m_i is the winner unit, $h_{ci}(t)$ the Gaussian neighborhood function, and $\alpha(t)$ the learning rate. Figure 2.5 depicts different levels of neighborhood for both square and hexagonal lattices. For example, there are 4 and 6 neighbors at level 1 in the square and hexagonal layouts, respectively, the hexagonal one providing a more uniform radial distribution. The basic idea in the SOM learning process is that, for each sample input vector P presented to the map/network for classification, the elements of the vector characterizing the winner neuron (or class) as well as those of the neighborhood nodes are updated to incorporate the new information, i.e., the elements of the winning and neighbor vectors in the map become closer to those of the input vector. In the early stages of training, the radius-defining neighborhood is large; and most of the SOM neurons strongly belong to any node's neighborhood. This creates an initial good global ordering of the SOM. As the training progresses, the radius is reduced to yield good local ordering as well.

Distances between map units reflect their topological relationships and are defined by the difference vectors between the prototype vectors representing the input samples respectively clustered in each one of them. The adjacent neurons or units of any neuron m_i define its neighborhood $N_i(t)$. Its value is generally defined as a decreasing function of the number of iterations or times (epochs) that the input data are presented to the map for training. On the other hand, the learning rate in equation (2.18) quantifies the fraction of the learning needed by each neuron at each training iteration. A complete description of the SOM and details of its implementation are provided in Kohonen (1995).

The SOM is both a powerful classification and visualization tool that provides several good alternatives for visualization purposes. The unified distance matrix, or U-matrix, is perhaps the most used method to display SOMs. U-matrix permits the visualization of the general cluster structure and the differentiation of each SOM cluster based on the distance between each codebook vector and its neighbors. In addition, each SOM node can be easily related to the input data space by using the component planes(C-planes). Straightforward correlations and relationships in the input data can be found by comparing several C-planes at the same time (Kaski, 1997; Kaski et al., 1999; Laine, 2003). Since SOM represents the similarity clustering of multivariate attributes, the visual representation becomes more accessible and easy to use for exploratory analysis. This

kind of spatial clustering facilitates the exploratory analysis of data with the purpose of identifying cause-effect relationships or correlations in exposure related problems when used in conjunction with environmental, transport and geophysical data.

2.4.3.1 Description of the SOM algorithm

i. Initialize the network: Set randomly the values of the initial weight vectors w_i at each node *i*. These weights are the elements of the prototypical vectors representing each neuron or cluster in the map. Set a large value to the initial neighborhood $N_i(0)$ to accelerate training initially.

ii. Present input data: Present each input pattern vectors *P(t)* to all nodes in the network at each iteration (epoch) *t*, either randomly or sequentially.

iii. Calculate the best matching unit (wining node): Calculate the node or neuron (BMU) characterized by the weights w_i that is nearest to the input pattern *P(t)* presented, as illustrated by Figure 2.6.

iv. Update weights: Update the weights of the prototype of the winning neuron and neighbor neurons with the following equation,

$$\begin{aligned} w_i(t+1) &= \{w_i(t) + h_{ci}(t)[P(t) - w_i(t)]\} \; if \; i \in N_c(t) \\ w_i(t+1) &= \{w_i(t)\} \; if \; i \notin N_c(t) \end{aligned} \tag{2.19}$$

where the term h_{ci} includes the neighborhood function and the learning rate and N_i represents the set of unit belonging to the neighborhood of the winner node.

v. Present the next input: Decrease h_{ci} so that $h_{ci}(t+1) < h_{ci}(t)$

Go to step ii and choose a new input vector *P(t)* until all input vectors have been presented at each iteration (epochs).

vi. Iterate until a predefined number of iterations or a threshold error is reached.

Usually, the quality of a SOM is evaluated by the mapping precision and its topology preservation capabilities. The *mapping precision* describes how accurately the neurons 'respond' to the given data set. It is represented by the quantization error, *qe*,

$$qe = \frac{1}{N}\sum_{i=1}^{N} \left\| x_i + m_c \right\| \tag{2.20}$$

where *x* is a sample vector of the input space and *m* is a reference vector representing any given unit within the map.

The topology preservation error, *te*, indicates how well the SOM preserves the topology of the data set,

$$te = \frac{1}{N}\sum_{i=1}^{N} u(x_i) \tag{2.21}$$

where $u(x_i)$ is 1 if the first and second BMUs or winning neuron (unit) are not next to each other, i.e., neighbors.

Lower quantization errors are associated to more accurate neuron responses in the SOM units. In addition, lower topographic errors imply better SOM topology preservation.

2.4.4 Recurrent neural networks

A recurrent neural network (RNN) is a ANN with feedback connections between neurons. This networks are computationally more powerful than other approaches such as feedforward neural networks. Recent applications include process control, speech recognition, handwriting recognition, adaptive robotics, music composition, protein analysis, stock market prediction, and many other sequence problems. Different types of RNN have been developed based on the degree of recurrent connections between the output neurons and the internal nodes: simple recurrent networks, partially recurrent networks and fully recurrent neural networks. This type of ANN was not used in this study and additional information can be found elsewhere (Mandic and Chambers, 2001).

2.4.5 Fuzzy ARTMAP networks

The fuzzy ARTMAP Neural Network (FAM) is a supervised learning mechanism capable of self-organizing stable recognition categories in response to arbitrary sequences of analogue and binary input patterns (Carpenter et al., 1992; Bartfai, 1995). It consist in a system constructed from two unsupervised ART networks (Adaptive Resonance Theory) (Carpenter and Grossberg, 1987).

There are two variations:

- ARTMAP which works with binary data
- Fuzzy ARTMAP which works with continuous data

The ART or fuzzy ART modules cluster the input and output patterns. During training ART_a receives input a^I and ART_b receives input b^I where b^I is the correct prediction (classification) for input a^I .

The two networks are linked using an associative learning network. This creates associations between the classes creates by ART_a and ART_b. The minimal number of ART_a categories needed for the desired accuracy is created.

It uses a minimax learning rule which minimizes error and maximizes code compression (the number of patterns stored in hidden units). The vigilance parameter, ρa, in ART_a is modified to control error. If a new category is needed the ρa is increased to force its creation. Lower values of ρa allow larger categories to form (greater compression) and when these are not appropriate then ρa is increased. The system modifies its parameters to reduce error. If ART_b does not predict the appropriate output for a given input to ART_a (predictive failure) then pa is increased by enough to force a new category to be created in ART_a. This operation is named match tracking.

Match tracking and *fast learning* allow the network to learn rare events which occur within a collection of frequent events which lead to different predictions. The network forces a category to be created for unique events so they can be identified as different. Figure 2.7 shows a schematic architecture of the fuzzy ARTMAP network.

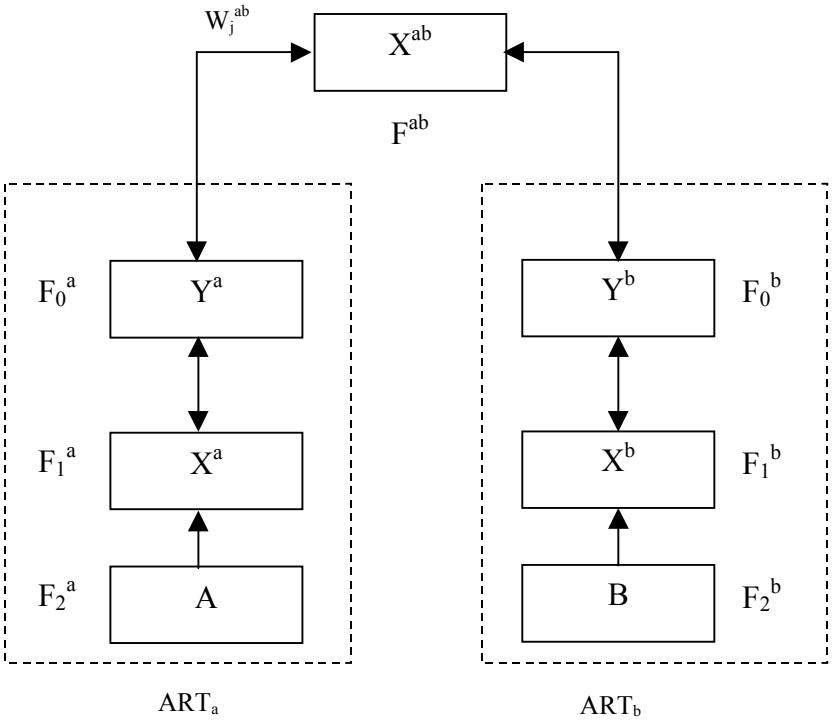


Figure 2.7. Fuzzy ARTMAP network architecture

2.5 References

ALLER, L., J. H. LEHR, R. PETTY and T. BENNETT (1987) "Drastic: A standardized system to evaluate groundwater pollution potential using hydrogeologic setting". United States Environmental Protection Agency. Project Summary EPA/600/S2-87/035.

BARTFAI, G. (1995) "An improved learning algorithm for the fuzzy artmap neural network". Victoria University of Wellington. Technical Report CS-TR-95/10. Wellington.

BOJÓRQUEZ-TAPIA L. A., G. M. CRUZ-BELLO, L. LUNA-GONZÁLEZ, L. JUÁREZ and M. A. ORTIZ-PÉREZ (2009) "V-DRASTIC: Using visualization to engage policymakers in groundwater vulnerability assessment". Journal of Hydrology 373(1-2): 242-255.

BURCHART, A., B. LEPPIG, A. MACDONALD, B. MÜLLER and G. WIMMER (2006) "Mapping the groundwater vulnerability in north rhine- westphalia, Germany". Environmental Engineering Science 23(4): 574-578.

CARLSSON C., A-K JOHANSON, G. ALVAN, K. BERGMAN and T. KÜHLER (2006) "Are pharmaceuticals potent environmental pollutants?: Part I: Environmental Risk Assessments of selected active pharmaceuticals ingredients". Science of the Total Environment 364(1-3): 67-87.

CARPENTER, G.A. and S. GROSSBERG (1987) "ART 2: Self-organization of stable category recognition codes for analog input patterns". Proceedings of the IEEE First International Conference on Neural Networks, San Diego, SOS Printing, San Diego, CA, USA: 727-735.

CARPENTER, G.A., S. GROSSBERG, N. MARKUZON, H. REYNOLDS and D.B. ROSEN (1992) "Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps." IEEE Transactions on Neural Networks 3: 698-713.

CHRISTAKOS, G. (2000) "Modern spatiotemporal geostatistics". Oxford Univ. Press, NY.

CHRISTAKOS, G., P. BOGAERT and M. SERRE (2002) "Temporal GIS". Springer, New York.

CHRISTAKOS, G., and M. SERRE (2000) "BME analysis of spatiotemporal particulate matter distribution in North Carolina". Atmospheric Environment 34: 3393-3406.

CIVITA, M. (1994) "Le carte della vulnerabilità degli acquiferi all inquinamento: Teoria and practica". Bologna, Pitagora Editrice.

DALY, D. and W. P. WARREN (1998) "Mapping groundwater vulnerability: The Irish perspective". Geological Society, London, Special Publications 130: 179-190.

DIXON, B. (2004) "Can an integrated ground water vulnerability mapping tool facilitate sensitivity analysis in a spatial domain?" Geo-Environment: Monitoring and Remediation of the Geological Environment: 151-160.

DIXON, B. (2005a) "Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: A gis-based sensitivity analysis". Journal of Hydrology 309(1-4): 17.

DIXON, B. (2005b) "Groundwater vulnerability mapping: A gis and fuzzy rule based integrated tool". Applied Geography 25(4): 327.

DOERFLIGER, N. and F. ZWAHLEN (1997) "Epik: A new method for outlining of protection areas in karstic environment". International Symposium on Karst Waters and Environmental Impacts, Antalaya, Turkey, Gunay and Jonshon.

DUIJVENBOODEN, W.V. and H.G.V. WAEGENINGH (1987) "Vulnerability of Soil and Groundwater to Pollutants". Proceedings of the International Conference on Vulnerability of Soil and Groundwater to Pollutants, Delft, The Netherlands.

E.E.A. (1998) "Environmental Risk Assessment – Approaches, Experiences and Information Sources". European Environmental Agency Publication.

ESCHER B.I., R. BAUMGARTNER, M. KOLLER, K. TREYER, J. LIENERT and C.S. MACARDELL (2011) "Environmental toxicology and risk assessment of pharmaceuticals from hospital wastewater". Water Research 45(1): 75-92.

ESCHER B.I. and K. FENNER (2011) "Recent Advances in Environmental Risk Assessment of Transformation Products". Environmental Science & Technology 45(9): 3855-3847.

FOSTER, S. (1987) "Fundamental concepts in aquifer vulnerability, pollution, risk and protection strategy". Vulnerability of Soil and Groundwater to Pollution, Proceedings and Information Nº 38. TNO Committee on Hydrological Research, Delft, The Netherlands, W. van Duijvanbooden and H. G. van Waegeningh.

GEMITZI, A., C. PETALAS, V.A. TSIHRINTZIS and V. PISINARAS(2006) "Assessment of groundwater vulnerability to pollution: A combination of gis, fuzzy logic and decision making techniques". Environmental Geology 49: 653-673.

GINEBREDA A., I. MUÑOZ, M. LÓPEZ DE ALDA, R. BRIX, J. LÓPEZ-DOVAL and D. BARCELÓ (2010) "Environmental Risk Assessment of pharmaceuticals in rivers: macroinvertebrate diversity indexes in the Llobregat River (NE Spain)". Environment International 36(2): 153-162.

GOLDSCHEIDER, N. (2005) "Karst groundwater vulnerability mapping: Application of a new method in the SwabianAlb, Germany". Hydrogeology Journal 13(4): 1431-2174.

HAGAN, M.T., H.B. DEMUTH and M.H. BEALE (1996) “Neural Network Design”. PWS Publishing, Boston, MA.

JONES R.N. (2001) “An Environmental Risk Assessment/Management Framework for Climate Change Impact Assessments”. Natural Hazards 23(2-3): 197-230.

KANGAS, J.A., T.K. KOHONEN and J.T. LAAKSONEN (1990) "Variants of self-organizing maps." IEEE Transactions on Neural Networks 1(1): 93-99.

KASKI, S. (1997) "Data exploration using self-organizing maps". Department of Computer Science and Technology. Dissertation for the degree of Doctor of Technology. Helsinki University of Technology.

KASKI, S., J. VENNA and T. KOHONEN (1999) "Coloring that reveals high dimensional structures in data." In T. Gedeon, P. Wong, S. Halgamuge, N. Kasabov, D. Nauck, and K. Fukushima, editors, Proceedings on ICONIP’99, 6th International Conference on Neural Information Processing II: 729-734.

KOHONEN, T. (1990) “The self-organizing map”. Neurocomputing 21: 1-6.

KRIGE, D.G. (1951) "A statistical approach to some basic mine evaluation problems on the witwateround." Journal of Chimic Mining Society 52: 119-139.

LAINE, S. (2003) "Using visualization, variable selection and feature extraction to learn from industrial data". Department of Computer Science and Technology. Dissertation for the degree of Doctor of Technology. Helsinki University of Technology.

LINDSTRÖM, R. (2005) "Groundwater vulnerability assessment using process-based models". PhD Dissertation. Royal Institute of Technology.

LAHR, J. and L. KOOISTRA (2009) “Environmental risk mapping of pollutants: State of the art and communication aspects”. Science of the Total Environment 408(18): 3899-3907.

LLOYD, C.D. and P.M. ATKINSON (2001) "Assessing uncertainty in estimates with ordinary and indicator kriging". Computers and Geosciences 27(8): 929-937.

LO J. (1998) “Multilayer perceptrons and radial basis functions are universal robust approximators”. IEEE International Conference on Neural Networks - Conference Proceedings 2: 1311-1314.

MANDIC, D. and CHAMBERS, J. (2001) “Recurrent Neural Networks for Prediction: Architectures, Learning algorithms and Stability”. Wiley.

MATHERON, G. (1963) "Principles of geostatistics". Economic Geology 58(8): 1246-1266.

MATHERON, G. (1971) "The theory of regionalized variables and its applications". France, Ecole des Mines, Fontainebleau.

MATHERON, G. (1973) "Intrinsic random functions and their applications". Advances in Applied Probability 5(3): 439-468.

MAO, Y.-Y., Z. XUE-GANG and W. LIAN-SHENG (2006) "Fuzzy pattern recognition method for assessing groundwater vulnerability to pollution in the Zhangji area". Journal of Zhejiang University Science A 7(11): 1917-1922.

MAZARI HIRIART, M., G. CRUZ BELLO, L.A. BOJÓRQUEZ TAPIA, L. JUÁREZ MARUSHI, G. ALCANTAR LÓPEZ, L.E. MARÍN and E. SOTO GALERA (2003)“Groundwater vulnerability assessment for organic compounds: Fuzzy multicriteria approach for Mexico City”. Environmental Management 37: 410-421.

MOORE and S. JOHN (1990) "Seepage: A system for early evaluation of the pollution potential of agricultural groundwater environments". Northeast Technical Center USDA. SCS, Geology Technical Note 5.

PANAGOPOULOS, G.P., A.K. ANTONAKOS and N.J. LAMBRAKIS (2006) "Optimization of the drastic method for groundwater vulnerability assessment via the use of simple statistical methods and GIS". Hydrogeology Journal 14: 894-911.

PARKER D.B. (1985) “Learning-logic: Casting the cortex of the human brain in silicon”. Center for Computational Research in Economics and Management science, MIT, Cambridge, MA.

RAHMAN A. (2008) “A GIS based DRASTIC model for assessing groundwater vulnerability in shallow aquifer in Aligarh, India”. Applied Geography 28(1): 32-53.

RALLO, R. (2007) “Multi-tier framework for the inferential measurement and data-driven modeling”.PhD dissertation.UniversitatRovira i Virgili. Spain.

RUMELHART D.E, G.E. HINTON and R.J. WILLIAMS (1986) “Learning representations by back propagating errors”. Nature 323: 533-536.

SAMPER, F.J. and J. CARRERA (1996) "Geoestadística. Aplicaciones a la hidrologíasubterránea". Barcelona, Centro Internacional de MétodosNuméricos en Ingeniería.

SANTOS J.L., I. APARICIO and E. ALONSO (2007) “Occurrence and risk assessment of pharmaceutically active compounds in wastewater treatment plants. A case study: Seville city (Spain)”. Environment International 33(4): 596-601.

SAVOLAINEN K., H. ALENIUS, H. NORPPA, L. PYLKKÄNEN, T. TUOMI and G. KASPER (2010) "Risk assessment of engineered nanomaterials and nanotechnologies – A review". Toxicology 269(2-3): 92-104.

SECUNDA, S., M.L. COLLIN and A.J. MELLOUL (1998) "Groundwater vulnerability assessment using a composite model combining DRASTIC with extensive agricultural land use in Israel's Sharon region". Journal of Environmental Management 54(1): 39-57.

SENAR E., S. SENER and A. DAVRAZ (2009) "Assessment of aquifer vulnerability based on GIS and DRASTIC methods: a case study of the Senirkent-Uluborlu Basin (Isparta, Turkey)". Hydrogeological Journal 17(8): 2023-2035.

SERRE, M. and G. CHRISTAKOS (2004) "Soft Data Space/Time Mapping of Coarse Particulate Matter Annual Arithmetic Average over the US." geoENV IV. Kluwer Academic Publishers, Dordrecht.

UMAR R., I. AHMED and F. ALAM (2009) "Mapping groundwater vulnerable zones using modified DRASTIC approach of an alluvial aquifer in parts of Central Ganga plain, western Uttar Pradesh". Journal of the Geological Society of India 73(2): 193-201.

URICCHIO, V.F., R. GIORDANO and N. LOPEZ (2004) "A fuzzy knowledge-based decision support system for groundwater pollution risk evaluation". Journal of Environmental Management 73: 189-197.

VAN STEMPOORT, D., L. EWERT and L. WASSENAAR (1993) "Aquifer vulnerability index: A gis compatible method for groundwater vulnerability mapping". Canadian Water Resources Journal 18: 25-37.

VÍAS, J. M., B. ANDREO, M.J, PERLES, F. CARRASCO, I. VADILLO and P. JIMÉNEZ (2006) "Proposed method for groundwater vulnerability mapping in carbonate (karstic) aquifers: The COP method". Hydrogeology Journal 14: 912-925.

VYAS V. and G. CHRISTAKOS (1997) "Spatiotemporal analysis and mapping of sulfate deposition data over the conterminous YSA". Atmospheric Environment 31(21): 3623-3633.

WORRALL, F. and T. BESIEN (2005) "The vulnerability of groundwater to pesticide contamination estimated directly from observations of presence or absence in wells". Journal of Hydrology 303(1-4): 92.

WORRALL, F. and D.W. KOLPIN (2003) "Direct assessment of groundwater vulnerability from single observations of multiple contaminants - art. No. 1345". Water Resources Research 39(12): 1345-1345.

Chapter 3

Groundwater Vulnerability Assessment

3.1 Introduction

Groundwater is a major resource of potable water for human population, and stake-holders are concerned about assessing and quantifying its quality and level of vulnerability.Industrialized and developed countries generate a high load of contaminants (stressors) that can damage the quality of its groundwater resources. Pollutants may reach groundwater usually by leaching through the ground and occasionally by direct contamination of wells. The fate of chemicals once they reach the soil depends on their physical and chemical properties, the degradation patterns and soil intrinsic characteristics (Mackay et al., 1996). The risk for groundwater contamination also involves the potential consequences of a contamination event (Lindström, 2005). Worrall et al. (2002) studied the interrelationships between the chemical properties of contaminants (e.g., solubility in water) and site properties (e.g., soil and aquifer type, and land use) using analysis of variance methodology (ANOVA). Their results provide statistical evidence that both chemical(or molecular) methods and site properties methods together (intrinsic vulnerability) have the potential to facilitate the understanding of groundwater pollution, as has been the case in pesticide contamination. Therisk of groundwater pollution has also been evaluated elsewhere from field measurements of contamination(Worrall and Kolpin, 2003; Worrall and Besien, 2005).

Intrinsic vulnerability maps define geographical areas that are “vulnerable” or sensible to pollution, independently of the pollutant characteristics.On the other hand, specific vulnerability assessment also considers stressors, such as the land uses, that are related tospecific contamination scenarios in the analysis (Martínez-Bastida et al., 2010). The availability of groundwater intrinsic vulnerability maps should help decision-makers (governments, national environmental agencies) in land planning for agricultural, industrial or urban use (Villa and McLeod, 2002; Passuello et al., 2012). Also, remediation plans can be initiated by national screening of groundwater resources using regional

vulnerability maps(Aller et al., 1987; Perles Roselló et al., 2009; Vías et al., 2010; Zabeo et al., 2011).

Monitoring groundwater quality is a common practice to address aquifers health. In many countries geo-referenced water quality data is publicly available and include measures for heavy metals, nitrates, nitrites and sometimes pesticides. The monitor strategy is fixed depending on the final use of groundwater resources, being the human consumption the most common use. Some legislative thresholds for potable water are established by local,regional and supra-national environmental administrations. Thresholds of selected pollutants in drinking water are shown in Table 3.1, according to theSpanish Real Decreto 140/2003, European Union Council Directive 98/83/EC and World Health Organization drinking water standards of 1993.

Table 3.1.Regulatory limits for pollutants in drinking water

Pollutant	Max. Threshold	Source
NO_2 (mg/l)	50	(a) (b) (c)
NO_3 (mg/l)	0.5	(a) (b)
SO_4 (mg/l)	250	(a) (b)
Fe (µg/l)	200	(a) (b)
Mn (µg/l)	50	(a) (b) (c)
Al (µg/l)	200	(a) (b) (c)
Sb (µg/l)	5	(a) (b) (c)
As (µg/l)	10	(a) (b) (c)
Ba (µg/l)	300	(c)
Cd (µg/l)	5	(a) (b)
Cu (µg/l)	2000	(a) (b) (c)
Cr (µg/l)	50	(a) (b) (c)
Mo (µg/l)	0.07	(c)
Ni (µg/l)	50	(a) (b)
Pb (µg/l)	10	(a) (b) (c)
Se (µg/l)	10	(a) (b) (c)
Zn (µg/l)	3000	(c)

(a) Spanish regulatory limit: Real Decreto 140/2003
(b) UE regulatory limit: Council Directive 98/83/EC
(c) World Health Organization: Drinking water standards 1993

Risk of contamination is thelikelihood or probabilitythatthe contaminant actually present in thegroundwater causes adverseeffects (end-points). LahrandKooistra (2009) presentedanoverview of the most important types of riskmaps, includingconcentrationmaps, exposuremapsandvulnerabilitymaps. Mapping thegroundwater quality information(concentration maps) provides a picture of the actual (and/or past) state of the aquifers due to the actual (and/or past) charge or presence of pollutantssourcesin each area. Regions with high pollutant's concentration can point out high vulnerable zones and/or intensive load of pollutants. On the other hand, areas with low or undetectable concentration can indicate either thatthezone has low vulnerabilityor thatthe load of pollutant over that specific area is low or null. Concentrations maps can

be used as validation for vulnerability maps following these rules: high pollutants concentration areas *must* occur in "high" vulnerability areas. On the other hand, consistent vulnerability maps should minimize the occurrence of high pollutant concentrations in"low" vulnerability areas.

Vulnerability maps have been used worldwide to assess the risk of groundwater contamination (Almasri, 2008; Andreo et al., 2005; Martínez-Bastida et al., 2010; Martínez-Santos et al., 2008; Masetti et al., 2009; Neukum et al., 2008; Rahman, 2008; Sinan and Razack, 2009). A literature search of existing vulnerability assessment schemes reveals that different models have been proposed in the past: Overlay and algebraic methods like DRASTIC (Aller et al., 1987; Burchart et al., 2006; Draoui et al., 2008; Goldscheider, 2005; Moore, 1990), statistical approaches (Panagopoulos et al., 2006; Worrall and Besien, 2005; Worrall and Kolpin, 2003; Nolan et al., 1997), fuzzy methods (Dixon, 2005a; Dixon, 2005b; Gemitzi et al., 2006; Mao et al., 2006; Mazari Hiriart et al., 2003; Uricchio et al., 2004), and process based simulation models (Lindström, 2005; Martínez-Santos et al., 2008; Nolan and Hitt, 2006; Popescu et al., 2008). First principle models should always be preferred but are difficult to apply to real conditions. On the other hand, overlay and algebraic methods have been developed to overcome the limitation of the large amount of monitoring data that statistical methods require, but at the cost of not providing information on the uncertainty of the vulnerability predictions.

Many authors have explored geostatistics techniques to generate smooth concentrations maps from measurement stations dataand comparison with different spatial interpolation methods (Tutmez and Hatipoglu, 2010; Kazemi and Hosseini, 2011). Hu et al. (2005) applied geostatistics techniques (kriging) to analyze the spatial variability of NO_3concentrations in groundwater.Also, different types of artificial neural networks have been applied as spatial interpolation methods. Zare et al. (2011) presented the advantages of the multilayer perceptron network for nitrate concentration mapping in comparison with linear regression methods. Self-organizing maps (SOM) (Kohonen, 1990; Kohonen, 2001; Vesanto and Alhoniemi, 2000)have been successfully applied in Environmental Risk Assessment for nitrate pollution in groundwater (Rallo, 2007). Rallo(2007) presented the interpolation and recovering of missing data capabilities of SOM for groundwater risk assessment of nitrate contamination. SOM were used to generate smooth concentrations maps by mimicking the kriging and co-kriging techniques from point source data. SOM have also been applied in many hydrogeological applications (Céréghino and Park, 2009; Kalteh et al., 2008; Peeters et al. 2007, Sánchez-Martos et al., 2002).

The aim of the current study is to apply SOM algorithms to classify and cluster intrinsic media variables to address groundwater vulnerability assessment in an integrated manner by defining and applying a new vulnerability index. Both spatial and temporal resolutionshave been considered. Spatial resolution has been examined at two different

scenarios, the local area of three counties known as the “Camp de Tarragona” andthe regional one of Catalonia. The temporal dimension has been accounted for by the analysis of a two-year period in the study area of Camp de Tarragona. The local scale approach has been used to demonstrate the capabilities of SOM to generate vulnerability maps by using the definition of a new vulnerability index. The regional scale study has been used to explore interpolation capabilities of SOM and to evaluate the behavior of the new vulnerability index during the up-scaling process. The classical DRASTIC Index analysis (Aller et al., 1987) was also performed in the study area to compare the proposed SOM-based vulnerability model and new vulnerability index with a standard and worldwide used methodology(Babiker et al., 2005).

Geostatistics by kriging interpolation has beenused to create smooth pollutant concentrations maps from measured georeferenced data of pollutants at the local scale assessment. Interpolation by SOM was used in the regional scale assessment to produce continuous maps of pollutants concentrations. The conjunction of pollutant exposure maps (concentration maps of exceeding legislative thresholds) of different contaminants allowedthe generation ofcumulative risk maps. Nitrates concentration and cumulative maps were used for the analysis and evaluation ofSOMvulnerability maps.

3.2 Area of study and data

The study is focused on Catalonia autonomous communitythat is located in the north-east of Spain (Figure 3.1). Catalonia is geologically diverse overitsarea of 32,107 km^2. Catalonia is formed by thefour administrative provinces of Barcelona, Tarragona, Girona and Lleida. Its capital and largest city is Barcelona, located in the Mediterranean coast. There are 41 counties and 946 municipalities that form independent local administrative units. The total population is approximately 7,360,000 (population census of 2008), which represents approximately the 16% of the total Spanish population. Near the 67% of this population is concentrated in the metropolitan area of Barcelona.

Catalonia has a very diverse orography, having a long coastline (547 km long), extensive mountain chains mirroring the coastline, mountains peaks reaching 3000 meters in altitude in the Pyrenees, inland depressions and the sedimentary delta of the Ebre river. Based on geological, hydrodynamic and geophysical properties, Catalonia is divided into 49 independent hydrogeological units, as shown in Figure 3.1a. The regional scale of the current study covers the whole area ofCatalonia. The Camp de Tarragona hydrogeological unit was selected for the local scale study(Figure 3.1b).

The climate is closely related to the diverse orography of Catalonia. In general, winters are mild (average temperature is 6-7 degrees Celsius) and summers are hot and dry (average temperature 24 degrees Celsius).Temperatures vary considerably between the coastline, the inland plains and the Pyrenees. In winter, the temperatures are considerably lower in the Central Depression and the Pyrenees, where thermometers can often measure temperatures several degrees Celsius below zero. At the coastline temperatures can reach more than 30 degrees Celsius in summer.

The geophysical characteristics of the Catalonia region were obtained from the Catalan and Spanish Governments. Hydrogeological maps were provided by the Catalan government,Departament de Medi Ambient i Habitatge and the InstitutCartografic de Catalunya (ICC), and the Spanish government, InstitutoGeológico y Minero de España (IGME) among others. Table 3.2 lists the sources, type of data and resolution of all available hydrogeological data used in the current study.

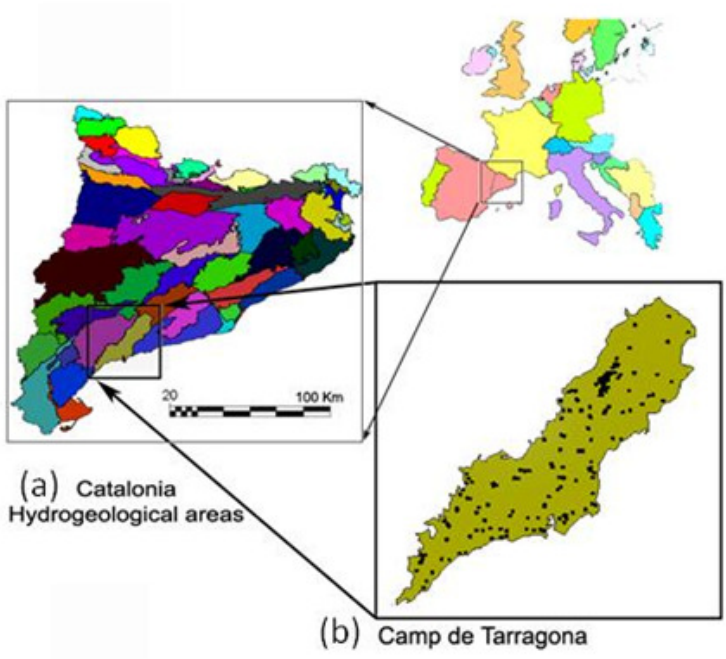


Figure 3.1.Spatial location of the area of study. (a) Regional scale: Catalonia. (b) Local scale: Camp de Tarragona

3.2.1 Regional scale: Catalonia

Available geo-referenced intrinsic media characteristics, like geological profiles, digital terrain model, annual rainfall maps, and land use for Catalonia, were processed by Geographical Information System (GIS) software to generate raster layers with a resolution 200 by 200 meters for the regional scale. The GIS MiraMon (Pons, 2006) was used for visualization and mapping. Figure 3.2 presents the digital terrain model of

Catalonia, where it can be observed the high variability in orography as described in the previous section. Figure 3.3 present the land uses map of Catalonia for year 2002. This figure shows the diversity of the types of land uses over Catalonia, where the urban and industrial areas are located mainly in the Barcelona area.

3.2.2 Local scale: Camp de Tarragona

The Camp de Tarragona forms a hydro-geological unit located in the south-east of Catalonia, close to the Mediterranean Sea (Figure 3.1b). This local area covers 406 km^2 and includes three counties (Tarragonès, Alt Camp and Baix Camp) which form independent administrative units. The area has a very dynamic economy with very important industrial, tourism related and agricultural activities in the context of Catalonia and the south of Europe. It includes two important cities, Tarragona and Reus, an airport and an industrial harbor. The total number of inhabitants is approximately 400,000.The economic activity is mainly industrial (41.6%) and services (43.3%), with real estate and construction (12%) and agriculture (3.1%) as the less important economic sectors.

Table 3.2. Sources and resolution of hydrogeological data

Data	Layer type (Original resolution)	Source
Geological	Raster (200 by 200 meters)	Departament de Medi Ambient i Habitatge
Land use	Raster (30 by 30 meters)	Departament de Medi Ambient i Habitatge
Annual rainfall	Raster (200 by 200 meters)	Departament de Medi Ambient i Habitatge
Digital terrain model	Raster (200 by 200 meters)	Institut Cartográfic de Catalunya (ICC)
Aquifer's permeability	Raster (1000x1000 meters)	Instituto Geológico y Minero de España (IGME)
Piezometrics level	Point data (559 measures)	Agencia Catalana del Agua (ACA)
Soil's permeability	Point data (123 measures)	Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas de España (CIEMAT).

The geo-referenced intrinsic media characteristics available, like geological profiles, digital terrain model, annual rainfall maps, annual media temperature, land use, and crop types, were processed with GIS (MiraMon) software to generate raster layers of 50 by 50 meters resolution. The following figures illustrate the GIS layer of Digital Terrain Model (Figure 3.4), Geological map (Figure 3.5) and land uses map (Figure 3.6) for the local Camp de Tarragona area.

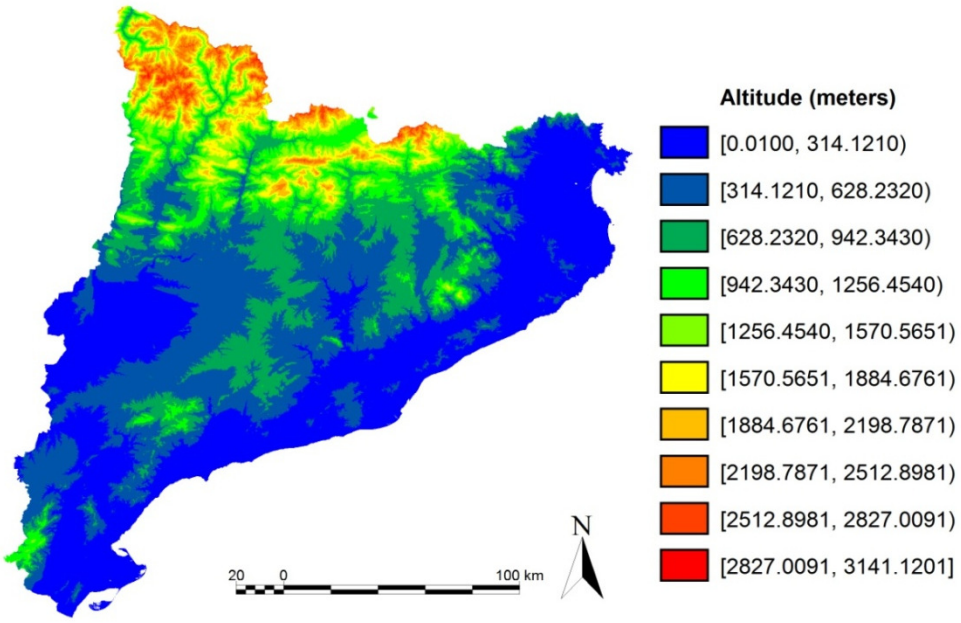


Figure 3.2. Catalonia Digital Terrain Model

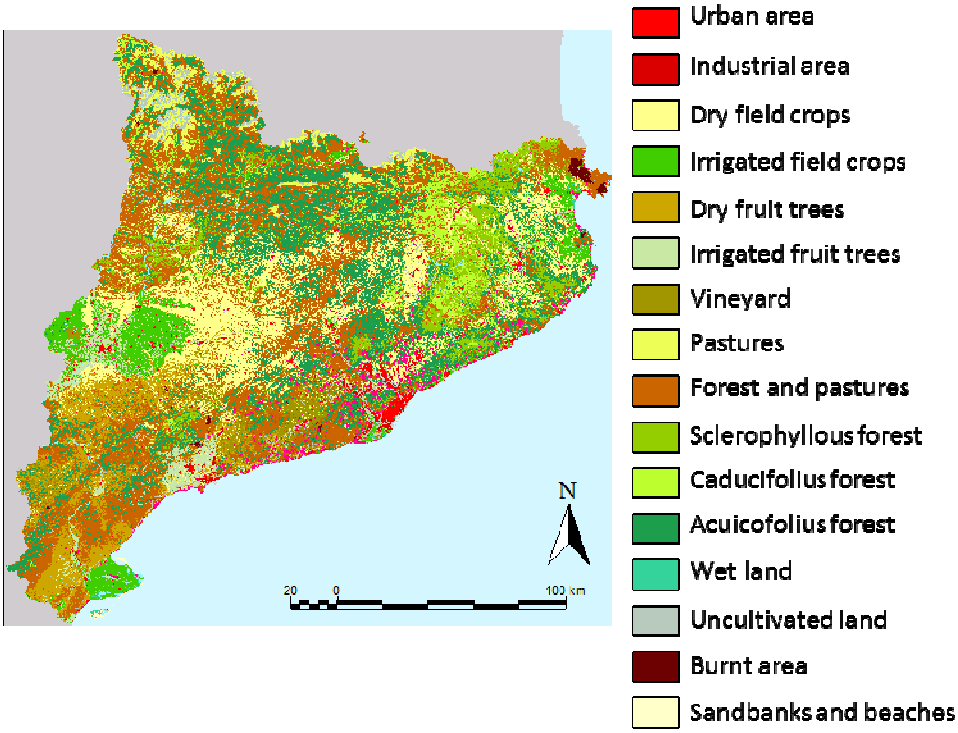


Figure 3.3. Catalonia Land uses map at year 2002

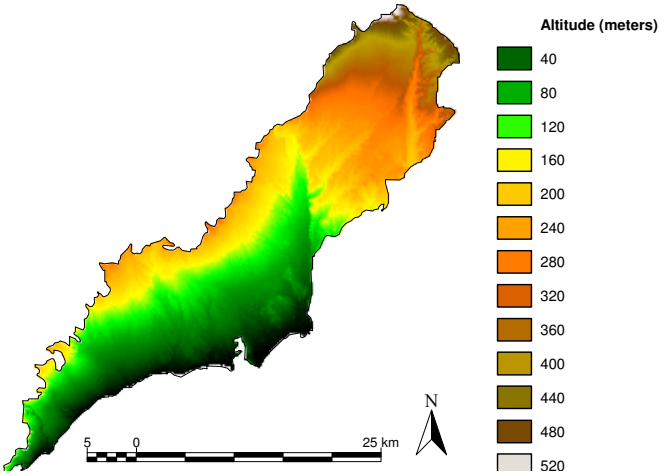


Figure 3.4. Camp de Tarragona Digital Terrain Model

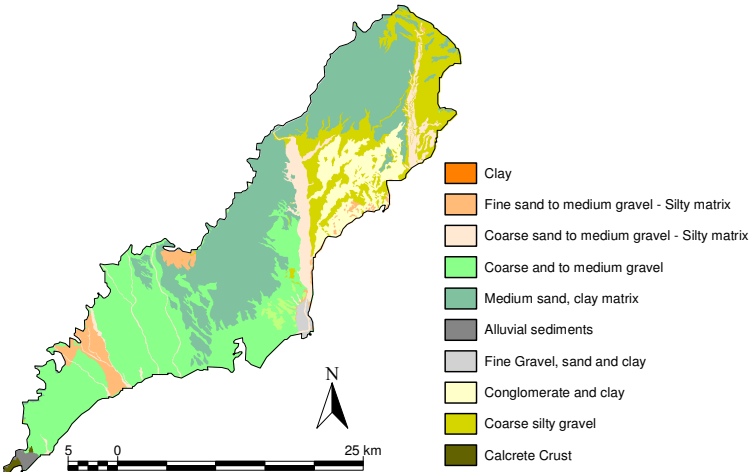


Figure 3.5. Camp de Tarragona Geological Map

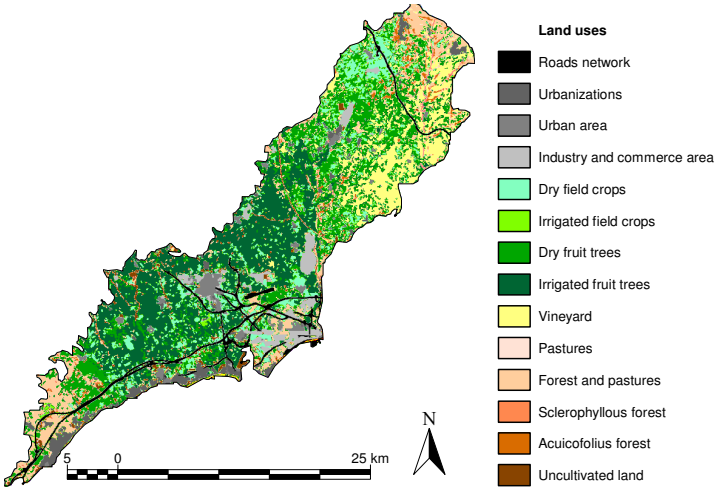


Figure 3.6. Camp de Tarragona Land uses map at year 2002

3.2.3 Pollution data

Groundwater data was provided by the Catalan Water Agency (ACA)for years 2002 and 2004. These data sets contain pollutant concentrations measured in 765 groundwater quality control points unevenly distributed over Catalonia (Figure 3.7); of which110 correspond to the Camp de Tarragona area. Data included heavy metals (Al, Sb, As, Cd, Cr, Mo, Ni, Fe, Mn, Se, Ba, Cu, Pb, Zn, and Mg), nitrate, nitrite and sulfates.

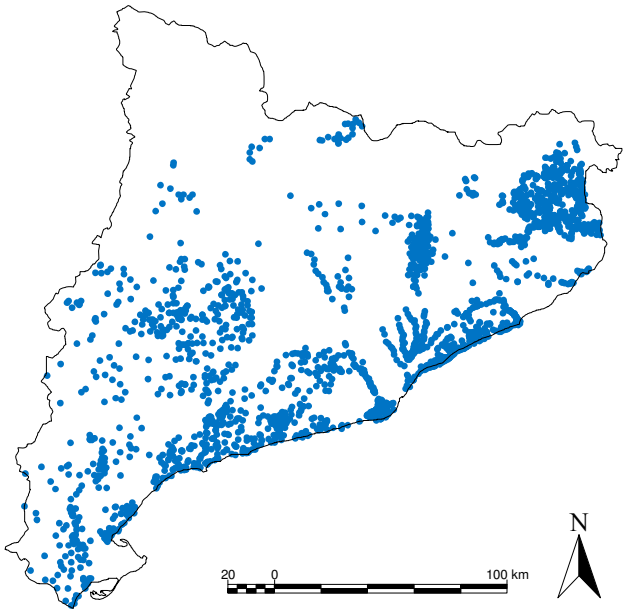


Figure 3.7. Groundwater quality control points in Catalonia area

Groundwater quality control network defined by Catalan Water Agency are subdivided into sub-networks comprised of wells from the same aquifer or territorial area delimited by hydrogeological, geographical or other criteria. The basic network is comprised by measurements points of physical-chemical composition of groundwater derived from water-environment interaction and diffuse pollution phenomena (not attributable to specific sources). Also a specific nitrate-vulnerable areas network was established by Catalan government and measures of majority of anions and cations, nitrogenous compounds (NO_3, NO_2, and NH_4) and metals formed in feed for animal consumption (Fe, Mn, Co, Zn, Se and Cu).

Tables 3.3 and 3.4 present a summary of the statistics of pollutants annual concentrations in groundwater for the areas under study at years 2002 and 2004. Data analysis shows that most pollutants exceeded legal thresholds (Table 3.1) in their maximum values or even in some cases in the mean values (for example, Fe, Mn and Mo). The same trend is observed for the two study areas. Only two pollutants (Cd and Cr) have concentrations values below legal limits for the period in study. Also, it can be noted that there is an

increase in nitrate mean concentration from year 2002 to 2004. This increase could be consequence of intensive load of nitrate to groundwater due to an increase of agricultural activities over Catalonia region.

Table 3.3. Statistics of heavy metals and pesticides in groundwater at year 2002

	Regional scale: Catalonia				Local scale: Camp de Tarragona			
	N	min	max	mean	N	min	max	mean
NO_2 (mg/l)	637	0.30	296.30	42.48	100	0.70	168.50	46.88
NO_3 (mg/l)	637	0.02	2.04	0.07	100	0.02	1.33	0.072
SO_4 (mg/l)	390	5.00	2704.00	198.00	68	30.50	496.00	138.62
Fe (µg/l)	401	20.00	82830.00	703.72	56	20.00	10680.00	599.2
Mn (µg/l)	401	5.00	5006.00	103.02	56	5.00	1048.00	79.57
Al (µg/l)	401	60.00	516.00	66.97	56	60.00	194.00	64.79
Sb (µg/l)	401	4.00	12.00	4.06	56	4.00	10.00	4.12
As (µg/l)	401	4.00	163.00	4.74	56	4.00	29.00	4.70
Ba (µg/l)	401	0.20	1020.00	81.61	56	11.7	1020.00	115.72
Cd (µg/l)	401	0.50	4.70	0.54	56	0.50	1.10	0.51
Cu (µg/l)	401	3.00	460.00	10.31	56	3.00	148.00	7.75
Cr (µg/l)	401	4.00	14.00	4.10	56	4.00	6.00	4.07
Mo (µg/l)	420	1.00	31.00	1.72	56	1.00	5.00	1.30
Ni (µg/l)	420	5.00	109.00	6.40	56	5.00	55.00	6.50
Pb (µg/l)	401	5.00	99.00	6.01	56	5.00	99.00	6.88
Se (µg/l)	401	3.00	108.00	10.20	56	3.00	22.00	6.91
Zn (µg/l)	401	4.00	7051.00	132.20	56	4.00	2600.00	129.84

N: number of measures; min: minimum value; max: maximum value; mean: arithmetic mean

Table 3.4. Statistics of heavy metals and pesticides in groundwater at year 2004

	Regional scale: Catalonia				Local scale: Camp de Tarragona			
	N	min	max	mean	N	min	max	mean
NO_2 (mg/l)	952	0.30	620.60	61.18	98	0.30	197.00	47.65
NO_3 (mg/l)	944	0.02	1.52	0.06	98	0.02	1.18	0.09
SO_4 (mg/l)	518	1.00	3552.00	243.58	57	14.00	630.00	149.18
Fe (µg/l)	505	20.00	37570.00	485.98	57	20.00	4021.00	276.68
Mn (µg/l)	505	5.00	3691.00	77.05	57	5.00	731.00	34.81
Al (µg/l)	505	60.00	7092.00	87.36	57	60.00	869.00	88.44
Sb (µg/l)	505	4.00	21.00	4.12	57	4.00	4.00	4.00
As (µg/l)	505	4.00	91.00	4.97	57	4.00	5.00	4.02
Ba (µg/l)	505	1.20	496.50	79.81	57	4.10	357.80	89.10
Cd (µg/l)	505	0.50	3.20	0.52	57	0.50	1.20	0.51
Cu (µg/l)	505	3.00	5001.00	19.43	57	3.00	296.00	11.95
Cr (µg/l)	505	4.00	23.00	4.85	57	4.00	23.00	4.49
Mo (µg/l)	505	1.00	246.00	3.08	57	1.00	17.00	1.46
Ni (µg/l)	505	5.00	119.00	6.85	57	5.00	77.00	6.74
Pb (µg/l)	505	5.00	95.00	5.89	57	5.00	25.00	6.63
(µg/l)	505	3.00	213.00	8.50	57	3.00	21.00	7.35
Zn (µg/l)	505	4.00	4597.00	119.84	57	4.00	1704.00	158.16

N: number of measures; min: minimum value; max: maximum value; mean: arithmetic mean

3.3 Local Scale Vulnerability Assessment: Camp de Tarragona

The capabilities of SOMin groundwater vulnerability assessment have been explored and a SOM-based vulnerability index is proposedatthe local scale study in the Camp de Tarragona area (Figure 3.1b). First, cumulative pollution maps have been generated using pollution data available over the area of study to evaluate the vulnerability maps obtainedwith DRASTIC andwith the new SOMindex as well. Second, DRASTIC vulnerability mapshave also been obtainedfollowing the methodology defined in Chapter 2. TheSOM capabilities to generate vulnerability maps have then been explored by using a DRASTIC-based SOMapproach and defining a new vulnerability index based on SOM clustering. Finally, a SOM-based vulnerability assessment was developed and validated using the cumulative pollution maps over the region of study. Matlab's SOM toolbox was used to perform self-organizing maps calculations (Vesanto et al., 2000) and principal functions are presented in Annex B.1.

The spatial coordinates of the area under study are defined by theUTM-31N-UB/ICC reference system(UTM: Universal Transverse Mercator coordinate system, UB: University of Barcelona, and ICC: Cartographic Institut of Catalonia) with an areal spam of 320500 to 367800meters in X-coordinate, and 4537500 to 4586000meters in Y-coordinate. The spatial resolution of all raster maps at the local scale developed in this study was 50 by 50 meters, generating a total of 917620 cells and 261479 active cells over the study region.

3.3.1 Cumulative pollution maps

The conjunction of pollutant exposure with exceeding legislative threshold values has permitted the assessment of cumulative risk for groundwater pollution and the generation of probability maps. Cumulative risk maps have been generated by addition of probability maps of those pollutants that exceed legal threshold. These maps have been used to asses both DRASTIC and the current new SOM vulnerability models.

3.3.1.1 Concentration maps

Geostatistics is used to create continuous surfaces from spatially distributed sample data. A theoretical model has been fitted to sample the variogram for each variable. Ordinary and Universal kriginghave been performed to generate concentration maps of the 17 pollutants of interest in a 50 by 50meters grid. The time-evolution of pollutants' concentrations has been obtained by applying this methodology over a two-year period and by comparing the potential exposure maps. To illustrate the procedure, Figure 3.8shows nitrates concentration maps for years 2002 and 2004. It is evident from this figure the increment of nitrate concentration with time and the presence of hot spots in the spatial concentration distribution.

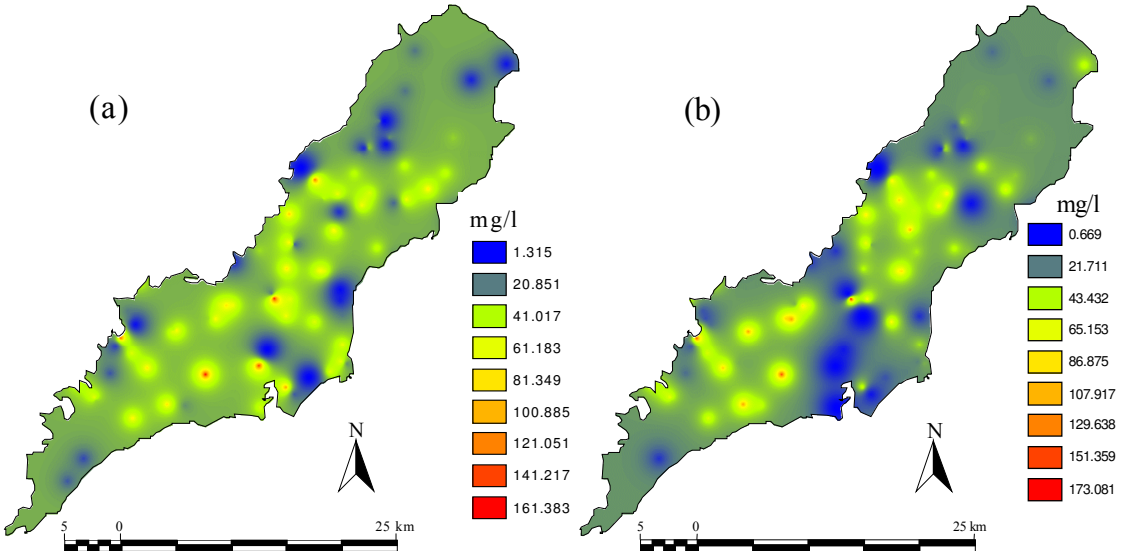


Figure 3.8. Spatial distribution of nitrates concentration generated by kriging interpolation in a two-year period. (a) year 2002, (b) year 2004

3.3.1.2 Cumulative exposure maps

The cumulative effects of pollutants can be studied by identifying elevated risk areas in cumulative exposure maps where high concentrations of different contaminants in water converge. Two types of maps are presented in this section. First, Boolean concentration maps have been created by assigning a value of one if the value exceeded legal threshold in the concentration maps and zero otherwise. The smooth cumulative exposure maps presented in Figure 3.9 have been generated by superposition of Boolean concentrations maps. The second approach is illustrated by the point cumulative maps presented in Figure 3.10.To build these maps, concentration data for each pollutant at each monitoring station are converted to Boolean data indicating if the annual value exceeds regulatory threshold (as indicated in Table 3.1). Afterwards, an intersection method is used to generate the cumulative maps for years 2002 and 2004, as shown in Figure 3.10.The principal differences in the two cumulative maps approaches arise from the data used to generate the maps and from the areal interpretation of the maps. In the smooth maps, concentrations maps generated by kriging were used, and some “hot spots” vanished due to the kriging interpolation technique. On the other hand, in the point cumulative maps the presence of “hot spots” is properly represented at each monitoring station. The smooth cumulative maps depicted in Figure 3.9 indicate that several one to more than five coincidences of pollutants exceeding legal threshold in the same geographical area occurred in the Camp de Tarragona area during the years 2002 and 2004. As expected, the highest concentrations of pollutants are found in the vicinity of the important industrial area and of the harbor, both located next to the city of Tarragona. These maps have been used to validate the current vulnerability approach.

Table 3.5. Frequency of annual pollutant concentration exceeding regulatory threshold at Camp de Tarragona at years 2002 and 2004

	2002	2004
NO_2	42	44
NO_3	3	1
SO_4	11	10
Fe	14	15
Mn	9	6
Al	0	4
Sb	1	0
As	2	0
Ba	3	1
Cd	0	0
Cu	0	0
Cr	0	0
Mo	56	57
Ni	1	2
Pb	1	6
Se	7	11
Zn	0	0
Total measures	119	109

Twelve and eleven pollutants exceeded in at least one monitoring station the legal threshold for year 2002 and 2004 respectively. Table 3.5 presents the frequency of annual pollutant concentration exceeding regulatory threshold at Camp de Tarragona used to generate point cumulative maps. These results indicate an increase accumulation of pollutant over the study area.

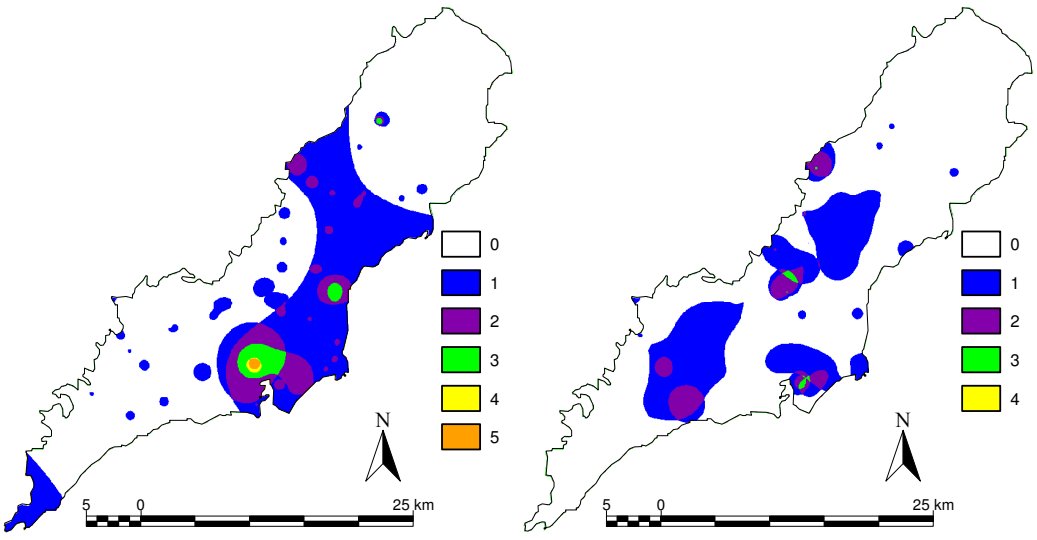


Figure 3.9. Smooth cumulative exposure maps. Combined effect of water pollutants exceeding regulatory thresholds. (left) Year 2002: Pb, Fe, Mn, Ba and nitrate. (right) Year 2004: Pb, Fe, Mn, Ba, Al, Se, nitrate and nitrite. The numbers in the labels indicate the number of pollutants exceeding legal threshold values

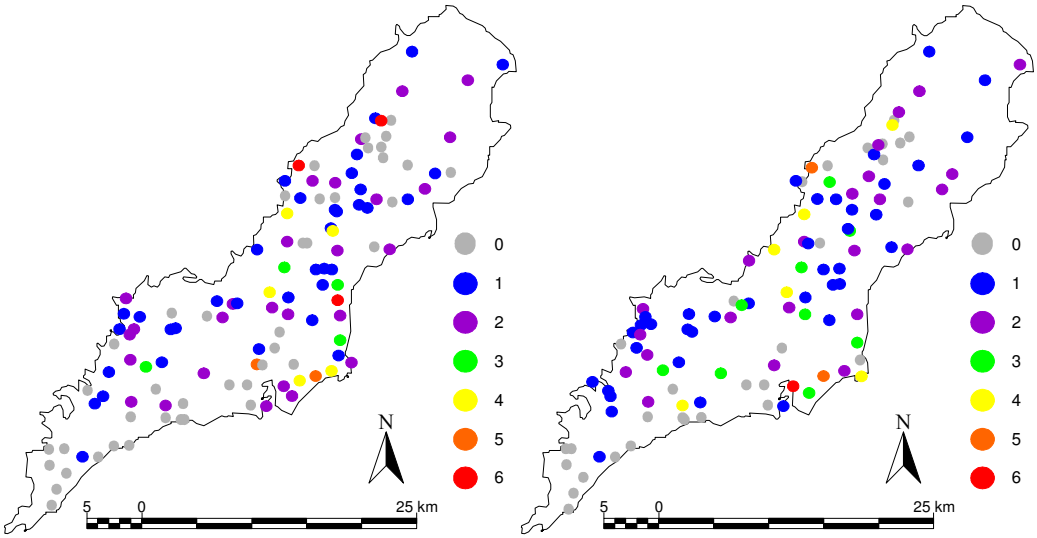


Figure 3.10. Point cumulative exposure maps. Combined effect of water pollutants exceeding regulatory thresholds (left) year 2002; (right) year 2004.The numbers in the labels indicate the number of pollutants exceeding legal threshold values

3.3.2 DRASTIC-based Intrinsic Vulnerability Map

In order to evaluate DRASTIC index and generate vulnerability maps each parameter in equation (2.1) has been ranked in a 1-10 rating according toPiscopo (2001) methodology, as described below. Table 3.2presents the available hydrogeological and climate data, type of data and resolution.

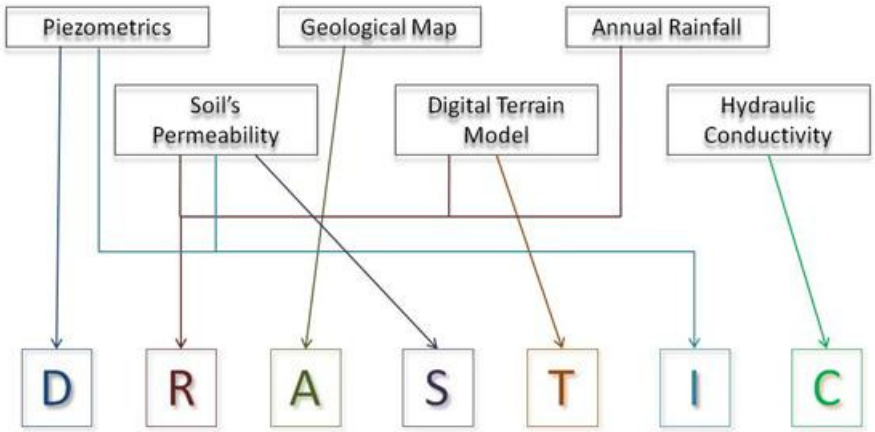


Figure 3.11. DRASTIC features generation from hydrogeological and climate data

Figure 3.11 summarizes the process for generating each DRASTIC feature from available data in the region of study. The DRASTIC index labeling of vulnerability areas was evaluated using different approaches published in the literature. Three labeling approaches were selected. The original DRASTIC labeling (Aller et al., 1987), that proposed by Draoui et al. (2008) to present a comparative study of vulnerability mapping

methods in a Mediterranean area, and the approach of Ahmed (2009) to highlight the differences between Generic and Pesticide DRASTIC models.

3.3.2.1 Depth to water (D)

Depth to water layer was generated from piezometrics data obtained from the Catalan Water Agency (ACA) and ConfederaciónHidrológicadel Ebro (CHEBRO).Rating categories are summarized in Table 3.6 and Figure 3.12.

Table 3.6. Depth to water rating for DRASTIC Index

Range (m)	Rating
0 – 5	10
5 –10	8
10 – 15	6
15 – 20	4
>20	1

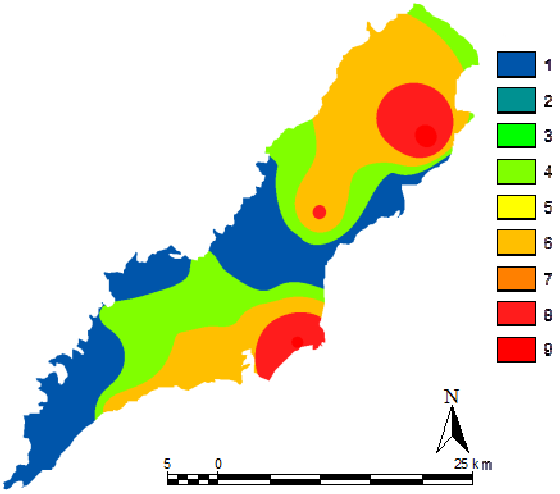


Figure 3.12. Depth to water layer for DRASTIC Index

3.3.2.2 Net recharge (R)

Based on data gathered from the Generalitat de Catalunya (GENCAT), net recharge parameter has been calculated by a linear combination of annual rain-fall, terrain's slope and soil's permeability (Piscopo, 2001).

Evapotranspiration has been reported constant within the whole Camp the Tarragona area. Equation 3.1 presents the calculation of Net Recharge feature, where %s, P and Ks are terrain's slope, annual rainfall and soil's permeability, respectively. They are rated as indicated in Table 3.7and rating categories for Rare shown in Table 3.8and Figure 3.13.

$$R = \%s + P + Ks \quad (3.1)$$

Table 3.7.Ratings for R calculation using equation 3.1

%s (%)	Factor	P (mm)	Factor	Ks	Factor
<2	4	>850	4	High	5
2-10	3	700-850	3	Mod-High	4
10-33	2	500-700	2	Mod	3
>33	1	<500	1	Slow	2
				Very slow	1

Table 3.8. Net Recharge rating for DRASTIC Index

Range	Rating
11-13	10
9-11	8
7-9	5
5-7	3
3-5	1

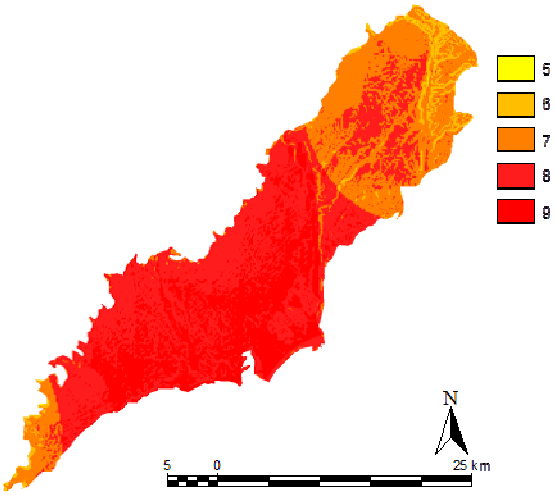


Figure 3.13. Net Recharge layer for DRASTIC Index

3.3.2.3 Aquifer media (A)

The geological map from GENCAT was processed to reduce the lithology categories. The ratting process has been performed using expert geological criteria. The lithology rating used for the DRASTIC index calculation is presented in Table 3.9. Aquifer media layer for Camp de Tarragona area is shown in Figure 3.14.

Table 3.9. Aquifer Media ratting for DRASTIC Index

Lithology type	Rating
Calcrete Crust	10
Coarse silty gravel	9
Conglomerate and clay	8
Fine Gravel, sand and clay	7
Alluvial sediments	6
Medium sand, clay matrix	5
Coarse sand to medium gravel	4
Coarse sand to medium gravel-Silty matrix	3
Fine sand to medium gravel - Silty matrix	2
Clay	1

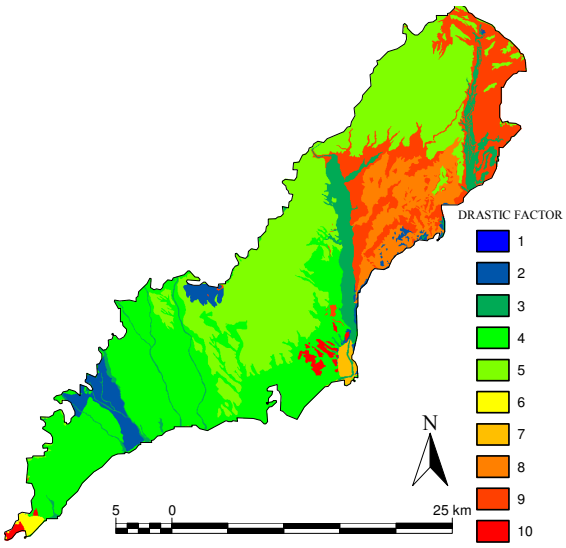


Figure 3.14. Aquifer Media layer for DRASTIC Index

3.3.2.4 Soil media (S)

The soil media map for the Camp de Tarragona was obtained from the soil's permeability map generated for Catalonia by kriging interpolation of soil's permeability point data provided by CIEMAT (Table 3.2).

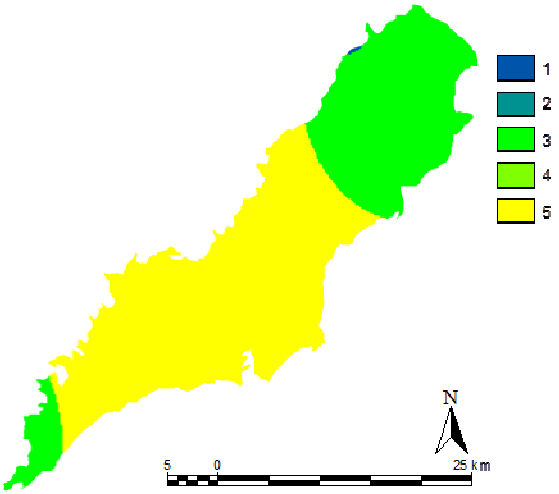


Figure 3.15.Soil Media layer for DRASTIC Index

3.3.2.5 Topography (T)

Slopes have been calculated from the Digital Terrain Model obtained from InstitutCartografic de Catalunya (ICC). Ratting has been performed based on the literature and expert criteria (Piscopo, 2001)(see Table 3.10).

Table 3.10. Topography ratting for DRASTIC Index

Range (%)	Rating
< 2	10
2 – 10	8
10 – 20	5
20 – 33	2
> 33	1

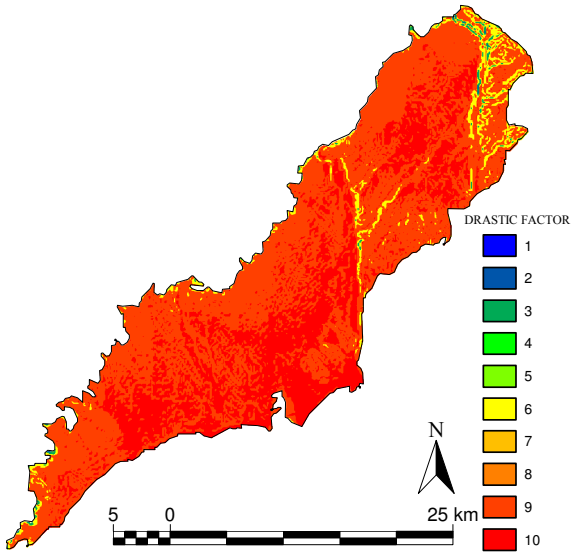


Figure 3.16. Topography layer for DRASTIC Index

3.3.2.6 Impact of vadose zone (I)

The type of vadose zone determines the attenuation characteristics of the material including the typical soil horizon and rock above the water table. The factors considered important in defining this parameter are soil permeability (or hydraulic conductivity) and depth to water table (Piscopo, 2001). The vadose zone impact is calculated by the following linear additive relation of soil permeability and depth to water table (equation 3.2) rated as indicated in Table 3.11. Final ratings for I feature are presented in Table 3.12 and Figure 3.17 shows areal distribution in Camp de Tarragona area.

$$\text{Vadose Zone} = \text{Soil Permeability} + \text{Depth to water table} \quad (3.2)$$

Table 3.11. Factors for Vadose Zone estimation

Soil Permeability	Factor	Depth to water (m)	Factor
High	5	0 – 5	5
Medium High	4	5 –10	4
Medium	3	10 – 15	3
Low	2	15 – 20	2
Very Low	1	>20	1

Table 3.12. Impact of vadose zone rating for DRASTIC Index

Range	Ratting
8 – 10	10
6 - 8	8
4 - 6	5
3 – 4	3
2 - 3	1

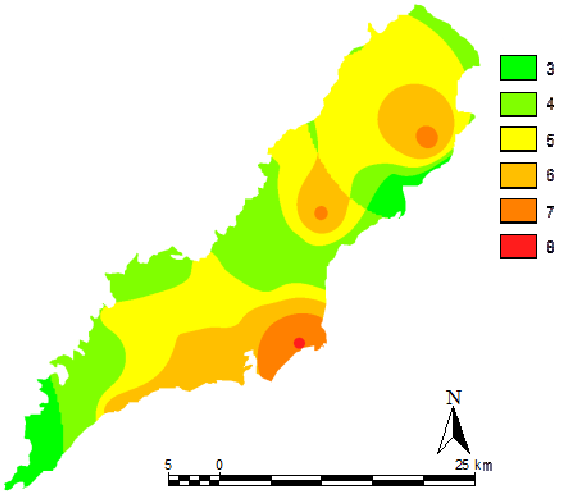


Figure 3.17. Impact of vadose zone ratting for DRASTIC Index

3.3.2.7 Hydraulic conductivity (C)

This parameter has been obtained from Instituto Geológico y Minero de España (IGME) from a raster map of 1000 x 1000 meters resolution for Spain. Catalan geologic department has been working on refining this information for the Catalonia area, but results are not finished by the time of developing this thesis.

Table 3.13. Hydraulic Conductivity rating for DRASTIC Index

Description	Rating
Very high conductivity – Not consolidated aquifers	10
High conductivity - Not consolidated aquifers	9
Very high conductivity – Karst aquifers	8
High conductivity – Karst aquifers	7
Low conductivity	3
Very low conductivity	2

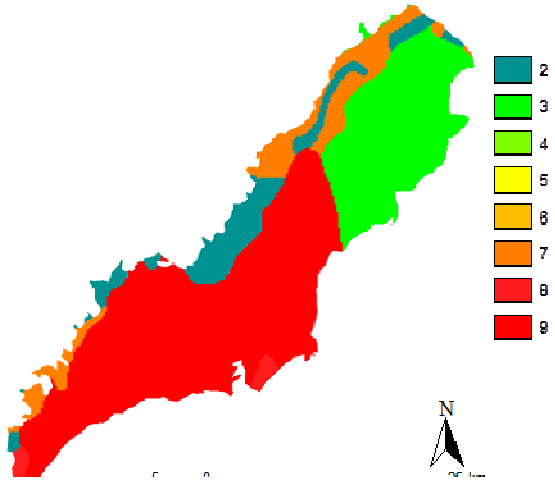


Figure 3.18. Hydraulic Conductivity layer for DRASTIC Index

3.3.2.8 DRASTIC maps

The DRASTIC vulnerability index was calculated for each grid cell following equation (3.3), i.e., by weighted addition of variable layers (generic DRASTIC weights in Table 2.2),

$$\text{DRASTIC Index} = 5D + 4R + 3A + 2S + T + 5I + 3C \quad (3.3)$$

As each DRASTIC feature is rated in a 1 to 10 scale, the resulting index will be within a range of 23 to 230.

The final step in the DRASTIC methodology is the labeling of the index or vulnerability class assignment. Three major vulnerability categories were defined in the original study of Aller et al. (1987), as presented in Table 3.14.Different labeling approaches in the DRASTIC index were studied in this work. Differences in vulnerability labeling are based on different percentile definition of the vulnerability index classes.

The Aller et al. (1987) labeling approach classifies the vulnerability index in three categories: low, medium and high vulnerability with percentiles cut of 48% and 68% (see Table 3.14). The Draoui et al. (2008) labeling defines five vulnerability classes: very low, low, moderate and high, with equi-espaced percentiles cuts of 20%, 40%, 60% and 80%, respectively (Table 3.15). In this case, the low and high classes were refined to account for more vulnerability differentiation. In the Ahmed (2009) labeling approach, only four vulnerability classes are defined: low, moderate, high and very high, with percentiles cuts of 50%, 65% and 75%, respectively (Table 3.16). The low and high percentile cuts in Ahmed approach are comparable to Aller labeling, with an emphasis in high vulnerability zonification.

The final use of the groundwater vulnerability map can determine the type of labeling that should be used to generate it (Lahr and Kooistra, 2009). Policy-makers, environmental agencies and local governments are interested in a global evaluation of groundwater vulnerability to generate actions plans, protective and/or remediation actions. At that policy level, vulnerability labeling using three major categories (low, moderate and high) is preferred for its simplicity. If remediation action plans are the goal, a more detailed vulnerability labeling should be adopted.

Table 3.14. DRASTIC vulnerability classes by (Aller et. al., 1987)

DRASTIC Index range	Color	Percentiles %	Vulnerability label
<79	Violet	0-48	Low
80-99	Indigo		
100-119	Blue		
120-139	Dark green	48-68	Medium
140-159	Light green		
160-179	Yellow	68-100	High
180-199	Orange		
>200	Red		

Table 3.15. DRASTIC vulnerability classes by (Draoui, et. al., 2008)

Generic DRASTIC Index range	Percentiles %	Vulnerability label
<63.6	0-20	Very low
63.6-104.2	20-40	Low
104.2-144.8	40-60	Moderate
144.8-185.4	60-80	High
>185.4	80-100	Very high

Table 3.16. DRASTIC vulnerability classes by (Ahmed, 2009)

Generic DRASTIC Index range	Percentiles %	Vulnerability label
<124	0-50	Low
124-150	50-65	Moderate
150-172	65-75	High
>172	75-100	Very High

Figure 3.19 shows the DRASTIC vulnerability maps obtained from the different labeling approaches. The original three DRASTIC vulnerability categories defined by Aller et al. (1987) and identified with the Index defined by Eq. (3.3) are shown in figure 3.19a for the Camp de Tarragona. Figure 3.19b depicts the three (moderate low, moderate and moderate high) out five vulnerability classes identified with the labeling approach of Draoui et al. (2008). Figure 3.19c illustrates the four vulnerability classes identified with labeling approach of Ahmed (2009). The visual inspection of these vulnerability and

DRASTIC features maps shows the high impact of C layer (hydraulic conductivity of the aquifer) in the low categorization of vulnerability.

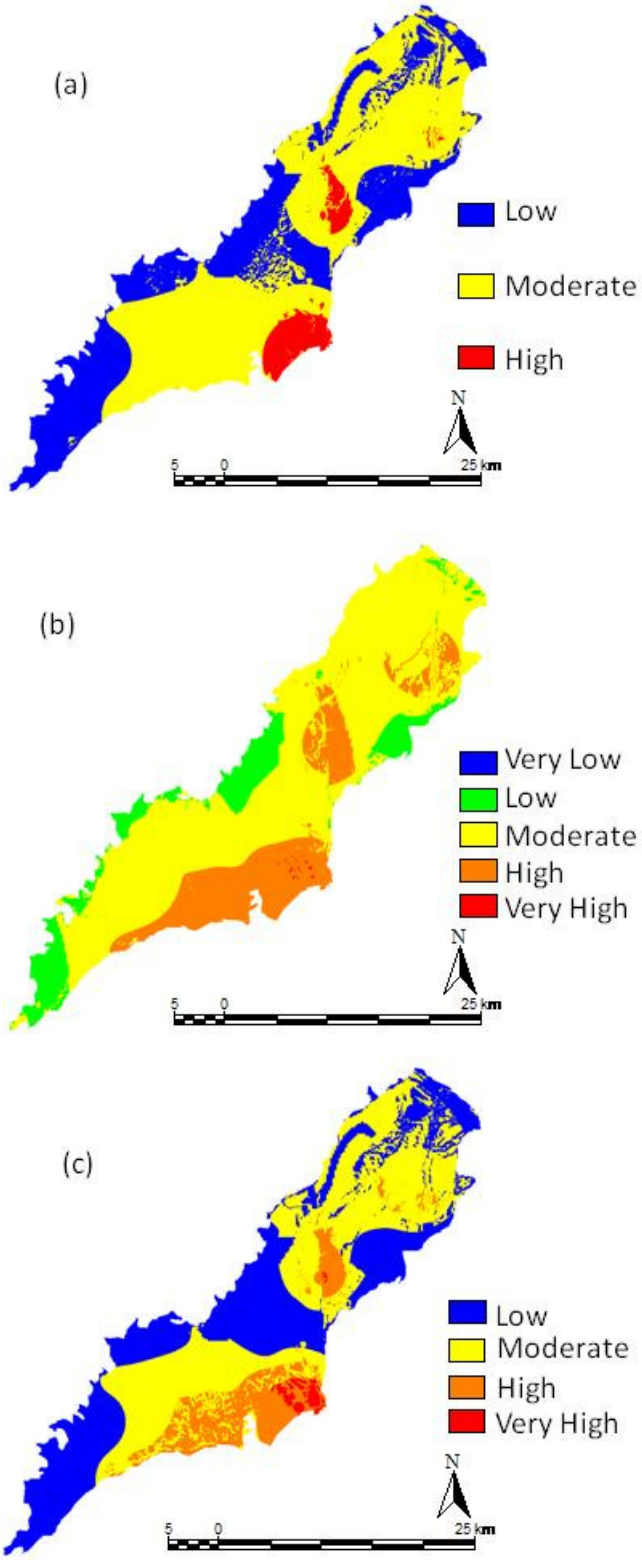


Figure 3.19. DRASTIC vulnerability maps for Camp de Tarragona at different vulnerability classes definitions: (a) Aller, (b) Draoui and (c) Ahmed

3.3.2.9 Assessment of DRASTIC vulnerability maps

The validation of a vulnerability map is not a simple task because there is not a direct measure of the vulnerability of a given area or region. Many authors have addressed the use of nitrate concentrations in groundwater as a measure of the vulnerability of an area (Stitger et al., 2006; Mishima et al., 2011; Tilahun and Merkel, 2010). Liggett and Allen (2011) evaluated the sensitivity of DRASTIC by using different data sources, interpretations and mapping approaches. They demonstrated that smaller-scale changes in vulnerability are not properly identified by the original DRASTIC methodology.

The vulnerability maps have been evaluated in the current study by using the concentrations of nitrates that exceed regulatory limits and also the cumulative effect of different pollutants exceeding legislative limits. The importance of a vulnerability map relies in the effective detection of high vulnerability areas that are of interest to environmental policy-makers for protection or remediation purposes. Areas with high nitrate concentration or cumulative pollutants' concentrations exceeding regulatory limits should be properly identified with the labels of moderate or high vulnerability by reliable vulnerability estimation approaches. In the other hand, areas with low nitrate concentrations and/or with non-exceeding regulatory limits should be related with either a low or null presence of pollutant sources or a low vulnerability of the aquifer.

As explained above, the current validation of vulnerability maps is focused on the proper identification of high vulnerability zones. Tables 3.17 to 3.19 present the frequency of any of the screened pollutants (Al, Sb, As, Cd, Cr, Mo, Ni, Fe, Mn, Se, Ba, Cu, Pb, Zn, Mg, nitrate, nitrite and sulfates) exceeding cumulative legal threshold in the Camp de Tarragona by year and vulnerability class for each of the three DRASTIC labeling approaches considered. Tables 3.20 to 3.22 present the same statistics for NO_3 only. The labeling approach of Aller et al. (1987) presented in Tables 3.17 and 3.20 reveals that the areas classified as "low vulnerability" include several measurement stations exceeding legal limits. In particular, Table 3.20 indicates that 10 and 11 points in the "low vulnerability" exceeded the limit for NO_3in 2002 and 2004, respectively. Also, the mean value of NO_3 concentrations is not consistent in the low, moderate and high vulnerability areas.

The corresponding analyses for the Draoui et al. (2008) and Ahmed (2009) labeling approaches yield similar results in Tables 3.18 and 3.19, respectively, since in their moderate-low and very low categories there are several cumulative points exceeding regulatory thresholds for different pollutants. Tables 3.21 and 3.22 showsame trends for exceeding NO_3concentrations in the low vulnerability areas for both labeling approaches.

Table 3.17. Frequency of exceeding cumulative legal threshold by year and vulnerability class in DRASTIC-Aller vulnerability map

	2002			2004		
#Exd	Low	Mod	High	Low	Mod	High
0	12	29	1	10	19	1
1	15	18	3	14	21	1
2	5	18	3	7	13	1
3	2	1	1	1	6	2
4	2	1	2	3	2	1
5	1	1	1	0	1	1
6	0	2	0	0	0	1

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.18. Frequency of exceeding cumulative legal threshold by year and vulnerability class in DRASTIC-Draoui vulnerability map

	2002					2004				
#Exd	Low	Mod Low	Mod	Mod High	High	Low	Mod Low	Mod	Mod High	High
0	-	7	22	13	-	-	5	17	8	-
1	-	3	25	8	-	-	4	26	6	-
2	-	2	17	7	-	-	2	14	5	-
3	-	0	3	1	-	-	0	5	4	-
4	-	1	1	3	-	-	2	3	1	-
5	-	0	0	2	-	-	0	1	1	-
6	-	0	3	0	-	-	0	0	1	-

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.19. Frequency of exceeding cumulative legal threshold by year and vulnerability class in DRASTIC-Ahmed vulnerability map

	2002					2004				
#Exd	Very low	Low	Mod	High	Very high	Very low	Low	Mod	High	Very high
0	12	-	22	8	0	10	-	15	4	1
1	16	-	16	3	1	16	-	18	2	0
2	8	-	13	4	1	8	-	11	1	1
3	2	-	1	1	0	2	-	4	3	0
4	2	-	1	1	1	3	-	2	0	1
5	0	-	0	2	0	0	-	1	1	0
6	1	-	2	0	0	0	-	0	1	0

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.20. DRASTIC-Aller statistics for NO_3 in Camp de Tarragona area for years 2002 and 2004

	2002			2004		
Stats	Low	Mod	High	Low	Mod	High
N	31	58	10	28	53	8
Nexd	10	25	6	11	27	2
mean	40.46	49.04	49.55	46.1	51.71	33.89
min	2.3	0.7	0.9	1.1	0.3	3.6
max	167.5	168.5	113	197	157.2	81.4
std	34.85	37.07	33.14	38.99	39.42	25.07
median	33.35	45.6	59.75	39.1	50.25	26.33
Q1	15.35	20.3	23.5	18.83	19.05	18.25
Q3	63.6	66.95	64.6	63.78	78.05	48.5

N: number of NO_3 measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

Table 3.21.DRASTIC-Draoui statistics for NO_3 in Camp de Tarragona

	2002					2004				
Stats	Low	Mod Low	Mod	Mod High	High	Low	Mod Low	Mod	Mod High	High
N	-	9	66	24	-	-	9	60	20	-
Nexd	-	2	25	14	-	-	4	26	10	-
mean	-	42.03	43.15	56.98	-	-	41.36	50.63	44.63	-
min	-	15.35	0.7	0.9	-	-	9.2	0.3	0.3	-
max	-	91.8	167.5	168.5	-	-	84.6	197	155.4	-
std	-	25.97	33.94	42.96	-	-	26.11	39.98	38.32	-
median	-	37.3	35.15	56.3	-	-	48.93	38.05	45.28	-
Q1	-	20.7	17.0	23.6	-	-	14.6	18.98	18.25	-
Q3	-	46.97	63.6	68.11	-	-	52.6	77.55	52.73	-

Symbols are the same as in Table 3.20.

Table 3.22. DRASTIC-Ahmed statistics for NO_3 in Camp de Tarragona area

	2002					2004				
Stats	Very low	Low	Mod	High	Very high	Very low	Low	Mod	High	Very high
N	35	-	49	12	3	32	-	45	9	3
Nexd	13	-	18	8	2	13	-	24	3	0
mean	42.0	-	43.98	68.21	50.2	45.28	-	53.82	39.63	24.95
min	2.3	-	0.7	0.9	23.5	0.3	-	0.3	0.3	22.2
max	167.5	-	141.8	168.5	64.6	197	-	157.2	155.4	30.1
std	34.45	-	30.48	55.22	23.15	38.2	-	36.21	52.1	4.46
median	35.5	-	38.7	60.5	62.5	39.1	-	50.5	14.3	22.55
Q1	15.35	-	20.3	20.25	23.5	18.03	-	28.2	3.6	22.2
Q3	67.05	-	60.1	106.35	64.6	64.63	-	79.85	56.7	30.1

Symbols are the same as in Table 3.20.

The above inconsistent results in Tables 3.17 to 3.22 for vulnerability categorization using DRASTIC-based approaches indicates that there is a need for a groundwater vulnerability approach that facilitates the integrated management of all georeferenced data and the calculation of a reliable vulnerability index for risk assessment. Self-organizing maps can provide the foundations for such an integrated approach.

3.3.3 SOM-based vulnerability map

The hypothesis behind the SOM-based vulnerability approach is that vulnerability assessment methods such as DRASTIC (Aller et al., 1987) could be mimicked by classification algorithms such as the self-organizing map (Kaski, 1997). The clustering capabilities of SOMhave been used to explore and identify homogeneous areas and relevant features that best discriminate each vulnerability zone in the physical map without using geographical coordinates. Two approaches have been developed to this end. The first consists of DRASTIC-based SOMmodel that usesthe same seven DRASTIC parameters for training the map (the parameters were rated as in the DRASTIC procedure) and estimates the intrinsic vulnerability. The second, is a specific vulnerability SOM model that self-organizes the former six (hydrogeological and climate) properties used to generate the DRASTIC features for the Camp de Tarragona area [piezometric level (H), annual rainfall (P), soil's permeability (Ks), land surface slopes (%s), aquifer media (A), hydraulic conductivity of the aquifer (Ka)] without the rating procedure for numerical variables, together with theland uses layer(land). This last layerhas been included in the vulnerability SOM model because it has been demonstrated to influence the vulnerability analysis (Secunda et al., 1998). It should be noted that the addition of the land uses layer implicitly incorporates specific stressor information* into the intrinsic vulnerability analysis(Chen et al., 2010). This addition converts any intrinsic vulnerability index, such as DRASTIC, into a specific vulnerability index to assess groundwater quality (Martinez-Bastida et al., 2010) by facilitating the specific relationship between the levels of contamination and the sources of potential contamination. In the following sections the specific vulnerability index [DRASTIC-Land uses] is implemented to evaluate local and regional groundwater contamination.

3.3.3.1 DRASTIC-based SOM model

To evaluate first the applicability of SOM in groundwater vulnerability assessment, a DRASTIC-based SOM model has been developed by self-organizing the seven DRASTIC features to generate vulnerability maps. The DRASTIC map developed in section 3.3.2 is

*Annual rain fall is considered as stressor in several ERA studies but not so for the intrinsic vulnerability analysis of groundwater

usedto assess the DRASTIC-basedSOMmodel. The input of this model is a vector with theseven DRASTIC features [depth to water table (D), net recharge (R), aquifer media (A), soil's permeability (S), topography (T), impact of the vadose zone (I) and hydraulic conductivity of the aquifer (C)] as elements.

A first result of this work, using SOM for clustering the seven DRASTIC features and DRASTIC weights to label the resulting vulnerability classes had been published as examples of the EU No miracle Project results (Pistocchi et al., 2011, see Annex A.1). Further work generated different scenarios and variables that have been studied to define finalSOM-based methodology for groundwater vulnerability mapping. Figure 3.20 shows a tree-graph of the different cases evaluated to study the capabilities of SOMfor groundwater vulnerability. The procedure followed is detailed below:

(i) The variables are clustered by a SOM and two scenarios are considered. One without variables weights (mask=1 for all variables) and a second one using DRASTIC weights (mask=DRASTIC weights normalized to have unitary sum);
(ii) A double clustering approach is considered with a second SOMdeveloped by using the output of the first SOM;
(iii) The possibility of calculating the vulnerability index using centroids of first or second cluster (cluster 1 and cluster 2 respectively) is explored. In the case of using cluster 1, the index is calculated by using centroids of the first SOM and a mean of the index for each cluster of second SOM is used to generate the final vulnerability index. In the cluster 2 case, the index is directly calculated from the centroids of the second SOM;
(iv) Two scenarios have been studied for the calculation of the vulnerability index. A new vulnerability index, *vIndex*, is defined according to

$$vIndex_j = \sqrt{\frac{\sum v_{ij}^{\,2}}{n}} \tag{3.4}$$

and a weighted index, *wIndex*, based on normalized DRASTIC weights, applied to the cluster's centroids

$$wIndex_j = \sum w_i v_{ij} \tag{3.5}$$

In these equationsv_{ij} is the value of the centroid of variable i of the cluster j and w_i is the weight of variable *I* (the summation of the weights is equal to one).

A total of 12 cases were generated and evaluated to define the final DRASTIC-based SOM methodology. Vulnerability categories which are independent of the spatial region considered can be statistically defined by assuming a normal distribution function for the values of the vulnerability index and using two percentiles (based on Aller labeling of DRASTIC index): 48 and 68%. This defines three vulnerability categories: low, moderate

and high. As discussed in section 3.3.2, the selection of vulnerability categories is still not clear in the groundwater vulnerability expert's community. Thus, the current proposal adopts the original three DRASTIC main categories as indicated in Table 3.23.

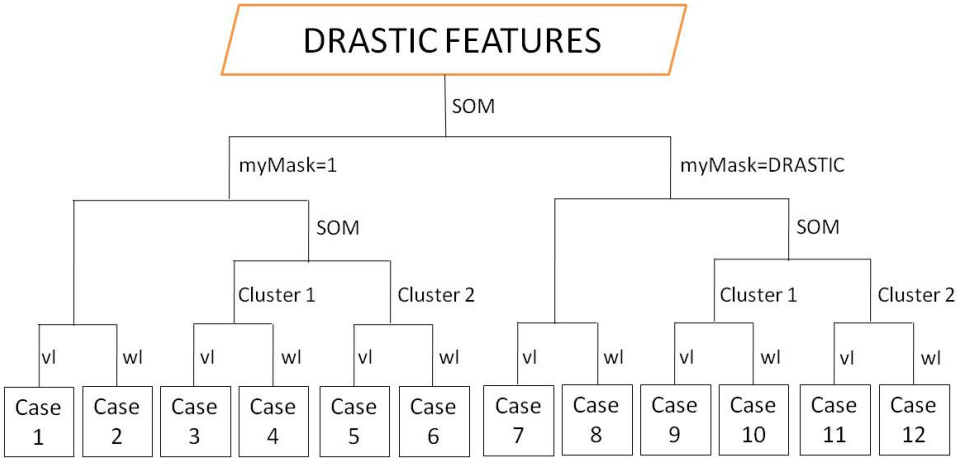


Figure 3.20. Case study definitions for DRASTIC-based SOMintrinsic vulnerability model. (vI: vIndex in equation 3.4 and wI: wIndex in equation 3.5)

Table 3.23. Vulnerability index categories

vIndex range	Vulnerability class
0-0.48	Low
0.48-0.68	Moderate
0.68-1	High

Based on the validation methodology presented in the previous section, that uses NO_3 concentrations and cumulative maps as a quality measure, case 11 in Figure 3.20was selected as the "best case" and defined as the final DRASTIC-based SOM vulnerability map. Figure 3.21presents the vulnerability map obtained following this double-SOM approach; normalized DRASTIC weights in the first SOM clustering and the *vIndex*calculation using the centroids of the second SOM. The SOM's size was selected by adopting different size configurations during training. A double-SOMwith 2542 and 260 units for the first and second map, respectively,wassufficient to extract vulnerability information from input hydrogeological variables without losing areal heterogeneities.

Table 3.24. Frequency of exceeding cumulative legal threshold by year and intrinsic vulnerability class in DRASTIC-based SOM vulnerability map

	2002			2004		
#Exd	Low	Mod	High	Low	Mod	High
0	0	30	12	0	23	7
1	2	26	8	2	28	6
2	1	13	12	1	12	8
3	0	2	2	0	3	6
4	0	2	3	0	4	2
5	0	0	2	0	1	1
6	0	3	0	0	0	1

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.25. DRASTIC-based SOM statistics for NO_3 in Camp de Tarragona area for years 2002 and 2004

	2002			2004		
Stats	Low	Mod	High	Low	Mod	High
N	3	66	30	3	60	26
Nexd	0	21	20	0	24	16
mean	24.28	40.67	61.24	25.78	45.9	56.58
min	18.8	0.7	0.9	13.8	0.3	0.3
max	33.25	167.5	168.5	48.93	197	155.4
std	7.91	33.3	38.88	20.06	37.39	40.85
median	20.7	33.8	63.25	14.6	35.58	51.8
Q1	18.8	16.85	36.4	13.8	18.85	22.55
Q3	33.35	60.1	78.2	48.93	67.18	86.1

Symbols are the same as in Table 3.20.

Statistics of NO_3 and cumulative exposure of pollutants for the DRASTIC-based SOM vulnerability model presented in Tables 3.24 and 3.25 show a significant improvement in the vulnerability classification compared to the previous DRASTIC results in Tables 3.17 to 3.22. Table 3.24 shows that only 3 measurements exceeding instances occur in the "low" vulnerability class for the complete set of pollutants.

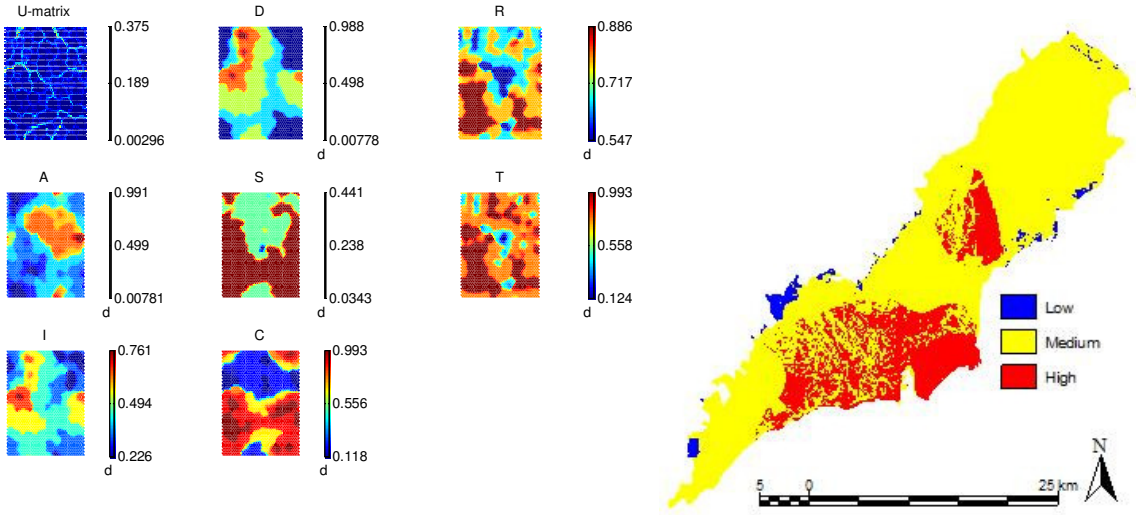


Figure 3.21. DRASTIC-based SOM intrinsic vulnerability model. (left) U-Matrix and c-planes for the seven DRASTIC features (D, depth to water; R, net recharge; A, aquifer media; S, soil media; T, topography; I, impact of vadose zone; C, hydraulic conductivity); (right) DRASTIC-based SOM vulnerability map for the Camp de Tarragona

As can be seen in Figure 3.21(right), the "low" vulnerability area predicted by the DRASTIC-based SOM model is very small and the NO_3 statistics indicates low concentration values in this area, mean 24.28 mg/l, without any concentration value exceeding the legal limit for NO_3 in Table 3.25. The correct identification of "moderate" and "high" vulnerability areas according to nitrates measures and cumulative point data reveals that the non-linear clustering capabilities of the SOM captures better the relationship between the seven DRASTIC features and vulnerability than it does the

simple aggregation method of the original DRASTIC approach. Using the self-organizing map to create clusters of similar patterns produce more reliable vulnerability maps than overlying methods such as DRASTIC.

Figure 3.21(left) depicts the U-matrix and the distribution of variables over the SOM space (C-planes) when the seven DRASTIC parameters are used to characterize the self-organization of classes within the Camp de Tarragona domain considered. These results are a necessary first step in the process of evaluating the performance of the current SOM procedure since they enable comparison with the DRASTIC vulnerability maps shown in Figure 3.19. The U-matrix shown in Figure 3.21(left) is capable of identifying vulnerability classes by the compactness of the clusters, i.e., by the small distances labeled with the blue color. Note that lighter blue, green and red areas identify larger distances and, thus, represent borders between classes.

An inspection of the C-planes in Figure 3.21(left) shows that the D and I features are correlated since their high (red) and low (blue) values have the same distribution over the map clusters. Thus, features D and I contribute in the same way to the map organization during the training process, i.e., they are redundant, and only one of the two should be considered.

3.3.3.2 SOM-based specific vulnerability model

The hypothesis that a SOM can mimic the DRASTIC methodology and generate reliable vulnerability maps using the same DRASTIC features has been demonstrated in the previous subsection. One step further, is the study of SOM capabilities to generate vulnerability maps avoiding some of the rating steps involved in the DRASTIC model. To evaluate this approach, a SOM-based vulnerability model has been developed using the raw hydrogeological data that generate the DRASTIC features, i. e., piezometric level (H), annual rainfall (P), soil's permeability (Ks), land surface slopes (%s), aquifer media (A), and hydraulic conductivity of aquifer (Ka). The inclusion of additional variables, such as land uses map (land), in the assessment has also been evaluated.

The vulnerability index defined by equation (3.4) has been used to generate the vulnerability categories defined in Table 3.23. The only requirement to compute the*vIndex* is that the numeric value of each variable considered contributes in the same direction, i.e., with a positive or a negative effect, and with the same proportion to vulnerability. All input variables are normalized in the range [0,1]before training the SOM. Table 3.26 lists the regional maximum and minimum values for each variable defined by expert criteria to quantify reliable vulnerability labels. Values in Table 3.26 assure the global applicability of the current vulnerability index. Regional maximum means the higher value that a parameter can reach in a regional sense and any value higher than the listed one represents the same vulnerability impact. On the other hand, regional

minimum means the lower value that a parameter can reach in a regional sense and any value lower than this represents the same vulnerability impact.

Numerical variables like piezometric level, annual rainfall, and land surface slopes were used with their raw values, without applying the expert criteria rating process used in the DRASTIC methodology. Categorical variables, like hydraulic conductivity, soil permeability and land uses, were rated as in the DRASTIC procedure.

Table 3.26. Regional maximum and minimum groundwater vulnerabilities for the SOM-based vulnerability features and analysis

Variable	Maximum	Minimum
Piezometric level (H)	50 m	0 m
Annual rainfall (P)	2000 mm	70 mm
Soil's permeability (Ks) *	10	1
Land surface slopes (%s)	60%	0%
Aquifer media (A) *	10	1
Hydraulic conductivity of aquifer (Ka) *	10	1
Soil Uses (land) *	10	1

* Categorical variables were considered using DRASTIC ratings

Figure 3.22(left) shows the U-matrix and the distribution of the seven selected variables over the SOM space (C-planes). The corresponding SOM vulnerability map is depicted in Figure 3.22(right). The topologies of the SOM structures were 2535 and 260 units forboththe first and second map. The quantization and topology errors for the final map were 0.038 and 0.076, respectively. The C-planes in Figure 3.22(left) reveal the independency of input variables (uncorrelated), as indicated by their different distributions over the output space in the trained SOM, where high and low values are respectively indicated by the red and blue colors.

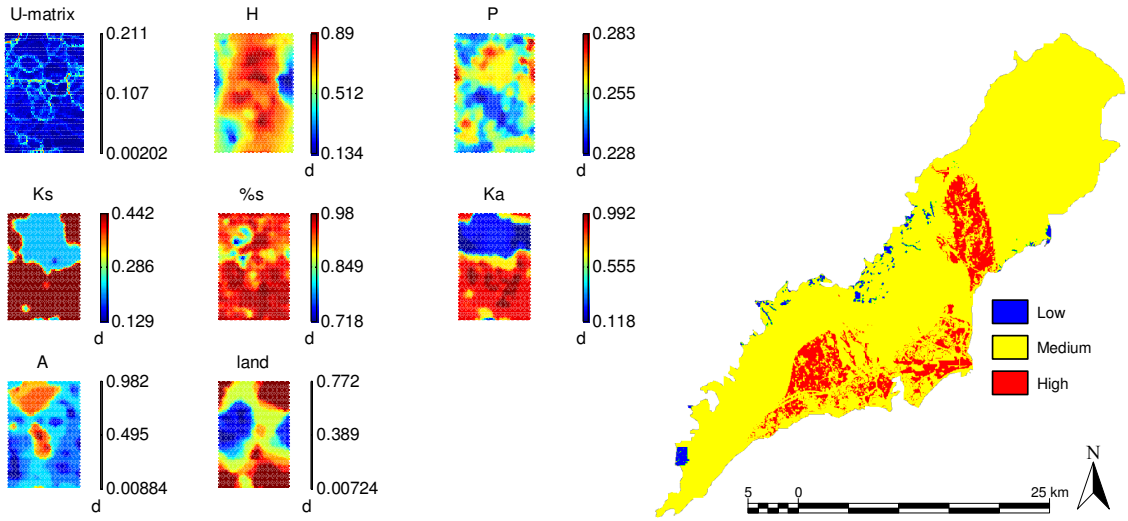


Figure 3.22. SOM-based specific vulnerability model (left) U-Matrix and C-planes for the seven input variables considered in the current SOM model (H, piezometric level; P, annual rainfall; Ks, soil permeability; %s, land surface slopes; Ka, hydraulic conductivity of the aquifer; land, land use); (right) SOMspecific vulnerability map for the Camp de Tarragona

Tables 3.27 and 3.28 present the statistics of the point cumulative map and nitrates concentration for the SOM-based vulnerability map of Figure 3.22. The “low” vulnerability zone is also small like in the DRASTIC-based SOM model (Figure 3.21).In this case, however, any monitoring stationis located in the “low” vulnerability areas identified by the proposed *vIndex*. For the “moderate” vulnerability area, approximately 40% of NO_3 measurement points exceeded the regulatory limit meanwhile at the “high” vulnerability area this value increases up to 57%. Obviously, the concentration distributions between these two vulnerable classes reflect any difference in NO_3 load distributions that they might have withstood.

The current SOM approach yields a specific vulnerability map that (i) provides a more detailed distinction between vulnerability zones within the area considered since they stem from the labeling of directly identified hydrogeological-climate classes, and (ii) correlates better with the cumulative map for combined effects of pollutants shown in Figures 3.9 and 3.10 than the previous DRASTIC and DRASTIC-based SOM maps presented in Figures 3.19 and 3.22(right), respectively. Both improvements arise from the use of a non-linear classification algorithm such as SOM to identify hydrogeological classes and to better relate them with vulnerability. Also, the adoption of a more appropriate set of parameters to categorize the domain and to assess vulnerability in the area considered are factors that improve vulnerability predictions. It should be noted that SOM preserves the topology of data and, even though geographical coordinates were omitted in the training process, the continuity of vulnerable areas (classes) in the physical domain is also maintained in the projection. In addition to the better discrimination and identification of vulnerable areas, the new SOM methodology can easily incorporate expert information other than DRASTIC in the groundwater vulnerability assessment.

SOM provides a good basis to select the most suitable set of variables for a specific area of concern since it effectively represents spatial regions of similar multivariable patterns which are identified and characterized by non-linear correlations between variables. The current approach minimizes the subjectivity involved in the DRASTIC rating process because the self-organization process that configures the map units into a topological representation of the original dataset is attained by means of an unsupervised training algorithm. The trained map nodes represent groups of entities with similar properties that reveal possible clusters (vulnerable areas) in the input data.

The SOM-based approach for groundwater quality (specific vulnerability) assessment using self-organizing maps yields reliable specific vulnerability maps from properly hydrogeological and climate variables, thus providing a sound basis for consistent land use planning and water resources management policies.

Table 3.27. Frequency of exceeding cumulative legal threshold by year and vulnerability class in SOM-based vulnerability map

	2002			2004		
#Exd	Low	Mod	High	Low	Mod	High
0	0	37	5	0	27	3
1	0	31	5	0	32	4
2	0	22	4	0	17	4
3	0	4	0	0	8	1
4	0	3	2	0	5	1
5	0	1	1	0	1	1
6	0	3	0	0	1	0

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.28. SOM-based specific vulnerability map statistics for NO_3 in Camp de Tarragona area for years 2002 and 2004

	2002			2004		
Stats	Low	Mod	High	Low	Mod	High
N	0	85	14	0	77	12
Nexd	-	33	8	-	34	6
mean	-	46.25	47.33	-	47.93	50.99
min	-	0.7	0.9	-	0.3	14.3
max	-	168.5	73.2	-	197	111.25
std	-	37.66	23.42	-	39.52	30.0
median	-	36.4	55.86	-	39.7	46.43
Q1	-	18.8	33.5	-	18.77	26.33
Q3	-	66.95	64.9	-	67.5	67.23

Symbols are the same as in Table 3.20.

TheSOM-based specific vulnerability approach developed and validated using cumulative maps and nitrates concentrations data at the local scale of the Camp de Tarragonauses expert criteria knowledge and physical information to define the maximum number of variables to be considered. The final methodology to generate SOM-based vulnerability maps is summarized below and illustrated in Figure 3.23.

(i) An initial set of variables should be defined by expert criteria and data availability of hydrogeological and climate data in the region under study.

(ii) Normalization of variables is needed as a preprocessing step. Regional maximum and minimum values for each variable have to be defined to assure standardization of final vulnerability maps as presented in Table 3.26.

(iii) Weights have to be defined for each proposed variable (DRASTIC weights can be used as expert criteria) using expert criteria and based on the physical impact of each variable in the vulnerability index. Weights have to be normalized[0,1] to be used as a *mask* in the training process of the first SOM.

(iv) Inspection of C-planes generated in the first SOM should help selecting the final set of independent variables to be used in the vulnerability index.

(v) A second SOM has to be trained with the final set of independent variables.

(vi) The vulnerability index defined by equation (3.4) can be calculated with the clusters centroids of the second trained map. Results of the second SOM are homogeneous clusters of the same vulnerability impact of georeferenced variables.

(vii) The final step is the labeling of clusters using the*vIndex* values of different clusters in three vulnerability categories (Table 3.23).

A peer-reviewed publication of SOM-based methodology for groundwater vulnerability assessment is presented in Mujica et al. (to be submitted) in Annex A.2 of this manuscript.

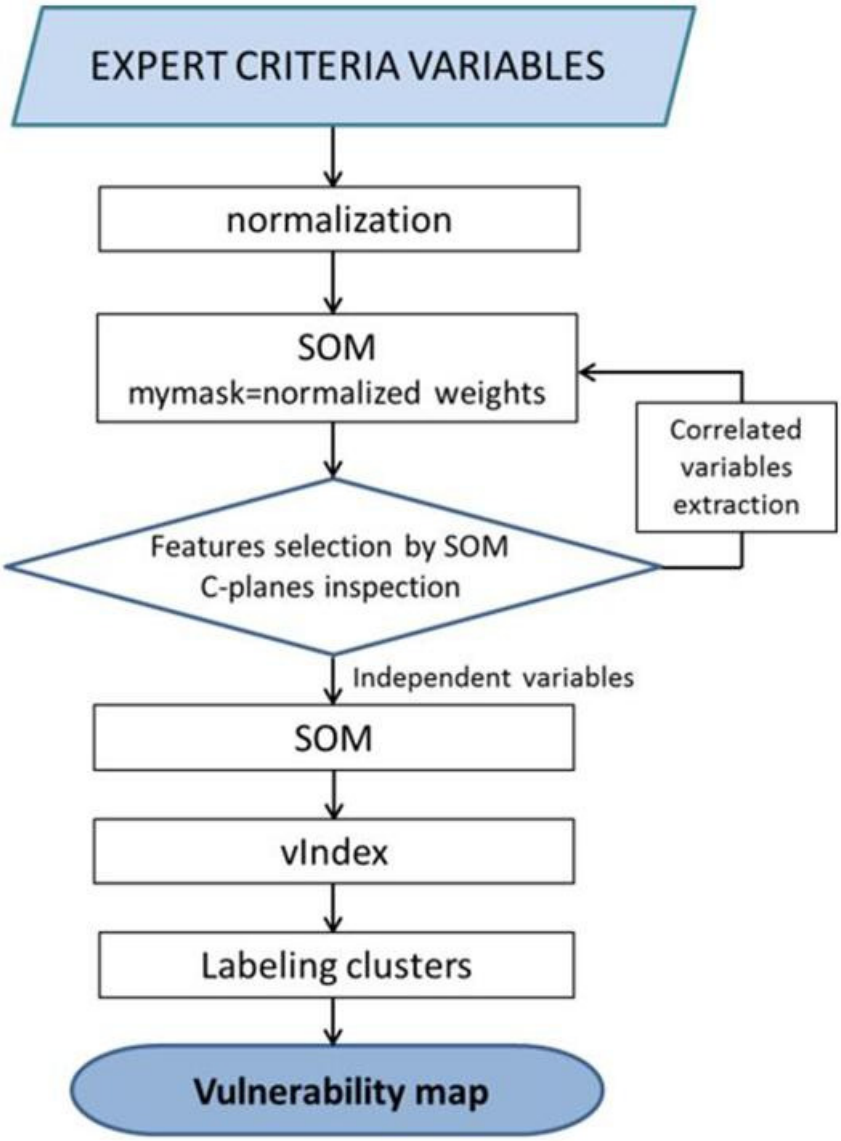


Figure 3.23. Scheme of the SOM-based specific vulnerability methodology

3.4 Vulnerability Assessment at Regional Scale: Catalonia

The SOM-based vulnerability methodology described in Figure 3.23 had been applied at the regional scale of Catalonia. While the number of measurement stations is usually enough at the local scale to yield accurate spatial interpolations, this is not always the case when considering wider regions, where some areas cannot be covered by the measurement network. To analyze data at the regional level an up-scaling process capable of spanning local data onto a wider region while maintaining spatial relationships is required. The intrinsic topology preservation properties of the SOM are exploited here to perform this up-scaling process and complete the development of the current integrated approach for vulnerability assessment.

The spatial coordinates of the area under study are defined by the UTM-31N-UB/ICC reference system (UTM: Universal Transverse Mercator coordinate system, UB: University of Barcelona, and ICC: Cartographic Institut of Catalonia) with an areal spam of 258000 to 526600 meters in X-coordinate, and 4485000 to 4752000 meters in Y-coordinate. The spatial resolution of all raster maps at the regional scale is200 by 200 meters, generating a total of 1792905 cells and 802739 active cells over the region of study.

3.4.1 Cumulative pollution maps

Two approaches are developed in this section. First, geostatistics is used to create continuous surfaces from spatial sample measurements; ordinary and simple krigingare performed to generate concentration maps of the pollutants presented in section 3.2.3. Second, the SOMis used to generate smooth maps by mimicking the co-kriging technique using hydrogeological units as constraints for the interpolation process (Rallo, 2007).

To illustrate the current analysis, Figure 3.24 shows nitrate concentration maps for year 2002 estimated using kriging interpolation and SOMinterpolation. The advantage of SOM interpolation as a cokriging technique for the mapping of pollutants' concentrations in aquifers is that it is geologically consistent due the fact that each hydrogeological area is an independent geological unit, i.e., there is no water flow continuity from one unit to the other. On the other hand, the consistent application of the kriging technique to each hydrogeological area requires enough geo-referenced data at each unit (statistically at least 10-15 points), which is impossible with the available data. The SOM-based method requires less data for producing smooth estimations of exposure concentrations.

The comparison of the resulting exposure concentration maps for nitrate shown in Figure 3.24 reveals that the results of the kriging interpolation (top figure) are mainly governed by the shape of the distribution model used to build the variogram. In contrast, the SOM-based interpolation (bottom figure) yields a less regular response due to the effect of the

inclusion of the spatial information concerning the hydrogeological units. In this later case, the interpolation process is mainly governed by the variables which constrain the modeled exposure. Even though the SOM-based interpolation yields higher estimates than the kriging one in Figure 3.24, the spatial locations of the main hot spots are preserved. Smoother exposure nitrate concentration distributions can be obtained by averaging the concentration values given by the SOM units located in the direct neighborhood of the best matching unit (BMU).

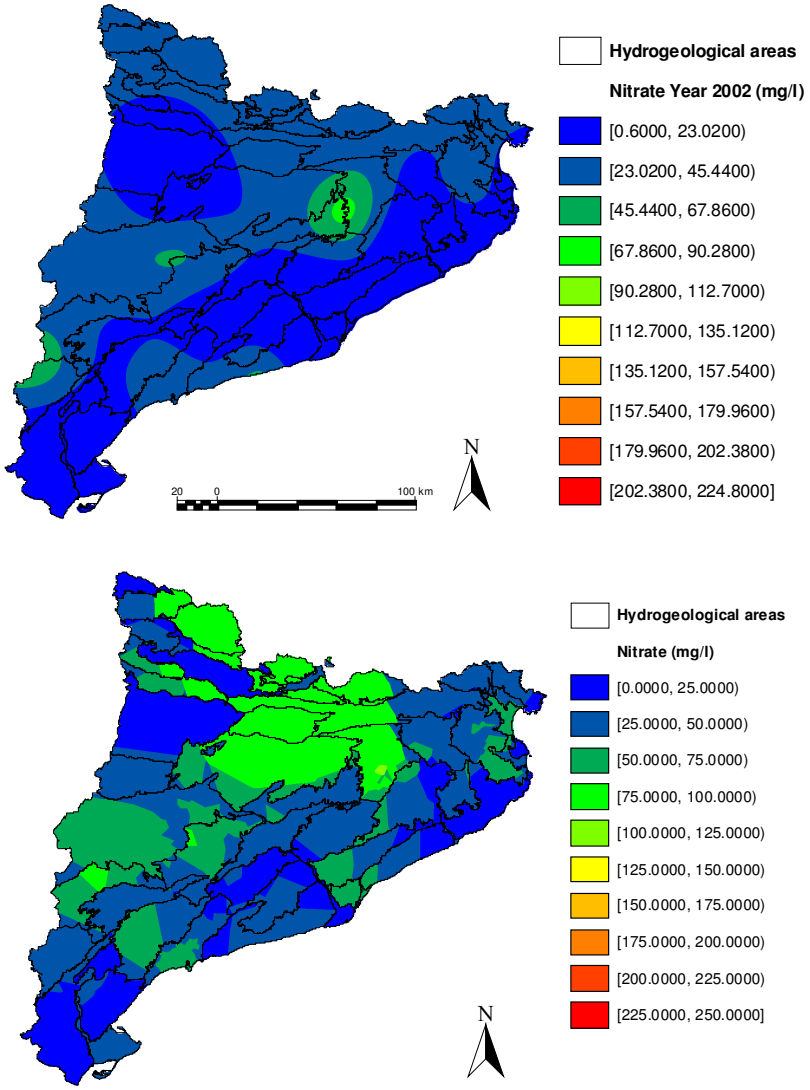


Figure 3.24. Spatial distribution of nitrate concentrations for year 2002 generated by (top) kriging interpolation (bottom) SOM interpolation

Exposure maps have been generated with both interpolation methodologies for the whole data set of 17 pollutants considered. Their cumulative effects have been studied with cumulative exposure maps to identify elevated risk areas where high concentrations of different contaminants in water converge. To build these maps the same procedure used for the local vulnerability analysis was followed, i.e., concentration maps were first converted to Boolean data (1's and 0's) indicating if a grid cell exceeded regulatory

threshold values or not, respectively, and afterwards, a Boolean intersection method for aggregation was used to generate the cumulative maps. Figure 3.25 depicts the resulting cumulative maps for year 2002 obtained using kriging and SOM interpolation techniques.

At year 2002, fourteen pollutants (NO_2, NO_3, SO_4, Fe, Mn, Al, Sb, As, Ba, Mo, Ni, Pb, Se, Zn) exceeded the legal threshold in at least one locationinCatalonia. Results range from 1 (Low) to 6 (High) in areas were one or more pollutants exceeded their legal threshold. It can be observed in Figure 3.25 that in almost all Catalonia at least one pollutant exceeded the legal threshold and a few hot spots of 6 pollutants exceeding legal threshold are also present.

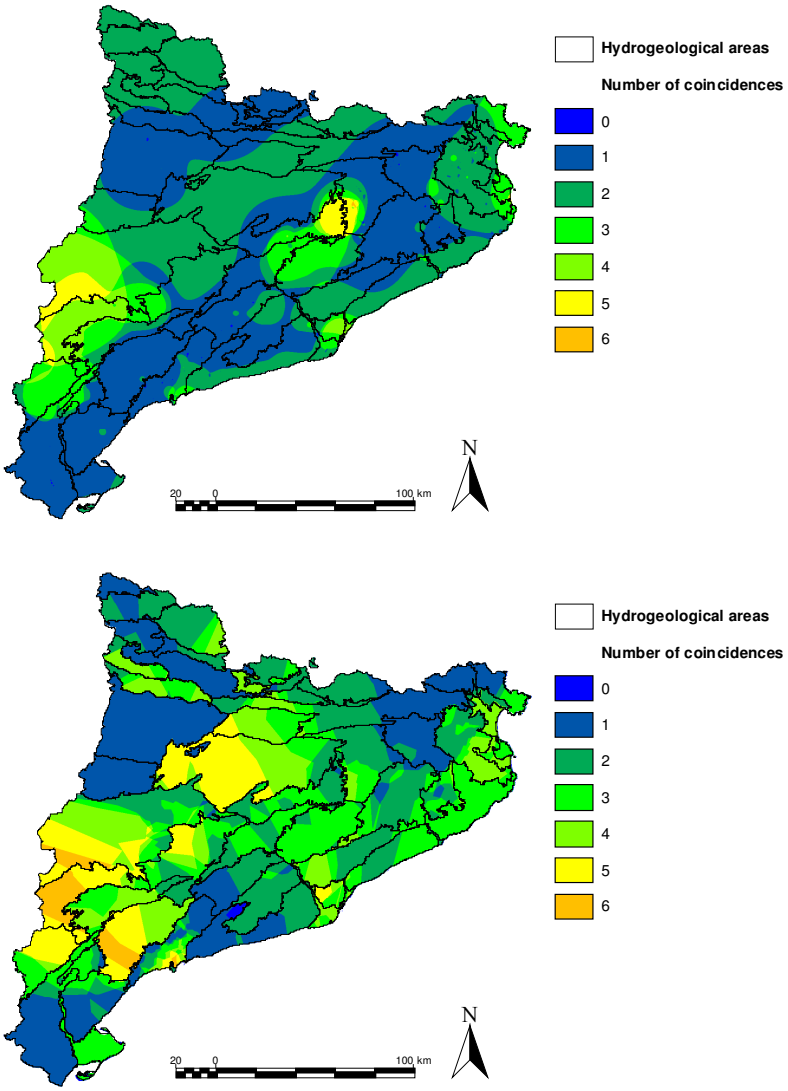


Figure 3.25. Cumulative exposure map for Catalonia in year 2002 generated by (top) kriging interpolation, and (bottom)SOM interpolation. Combined effect of water pollutants exceeding regulatory thresholds (Pb, Fe, Mn, Se, sulfate and nitrate).The numbers in the labels indicate the number of pollutants exceeding legal threshold values

Inspection of these cumulative maps shows again that the response of the kriging-based maps is mainly driven by the distribution function model assumed for the variogram. Again the spatial location of hotspots is consistent in these two map representations. The cumulative map obtained by the combination of single pollutant interpolated maps using SOM gives slightly higher estimates than the one obtained using kriging. This is mainly due to the inclusion of the continuity in the hydrogeological units which spans the pollutant concentrations to nearby locations belonging to the same hydrogeological unit. These results are consistent from the point of view of mechanistic transport of pollutants in groundwater. The kriging approach cannot reproduce these results because it is dependent on the spatial locations of the measurement stations for which the pollutant concentrations are known. In should be noted that the exposure maps produced using SOM interpolation can be optimally smoothed by using the neighbor averaging technique discussed previously.

3.4.2 DRASTIC vulnerability model

Vulnerability maps have been generated with equation 3.3 to evaluate the DRASTIC index, and have been used as a reference to evaluate the current SOM approach at the regional scale of Catalonia. As explained previously the equation parameters were ranked with a 1-10 rating according to the DRASTIC expert criteria. The same methodology explained in section 3.3.2 for the local scale was performed at the regional scale to calculate DRASTIC features (Piscopo, 2001). Depth to water table (D) was generated from piezometric data using the ratings presented in Table 3.6. Net recharge layer (R) was obtained from the addition of surface slopes, annual rainfall and soil´s permeability layers, and rated as presented in Tables 3.7 and 3.8. The aquifer media (A) and hydraulic conductivity (C) features were generated from aquifer permeability data rated by the definitions in Table 3.13. Soil media layer (S) was obtained from infiltration capacity data. Topography (T) was generated using the DTMrated using Table 3.10. Impact to vadose zone layer (I) was generated from the addition of soil`s permeability and the depth to water table layer, as stated in Tables 3.11 and 3.12.

The DRASTIC vulnerability map for Catalonia shown in Figure 3.26was computed with three categories, ranging from Low to High vulnerability areas. It can be seen in this figure that the most highly vulnerable areas are located in the east part of Catalonia near the MediterraneanSea. Tables 3.29 and 3.30 present pollutants statistics for the vulnerability areas identified by the DRASTIC methodology over Catalonia. Nitrates pollution data reveal that “low” vulnerability class has higher mean concentration value than “moderate” and “high” zones (47.73 mg/l for the “low” class but 40.02 and 35.22 mg/l for “moderate” and “high” classes). Also, the “low” vulnerable class includes the 47% of monitoring stations exceeding regulatory limits for all pollutants considered in the study. These results highlight again at the regional scale of Catalonia the drawbacks of the

DRASTIC methodology already observed and documented at the local scales of the Camp de Tarragona.

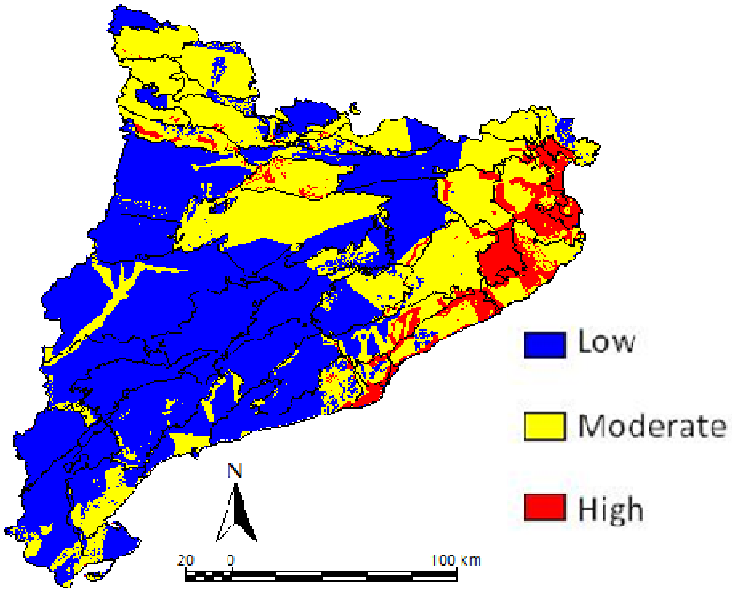


Figure 3.26.DRASTIC vulnerability map for Catalonia

Table 3.29. Frequency of exceeding cumulative legal threshold for each DRASTIC vulnerability class for Catalonia in 2002

#Exd	Low	Mod	High
0	126	55	57
1	128	62	32
2	50	53	26
3	35	26	33
4	16	11	17
5	5	2	8
6	3	2	0

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.30. DRASTIC vulnerability map statistics for NO_3inCatalonia at year 2002

Stats	Low	Mod	High
N	317	172	148
Nexd	117	50	41
mean	47.73	40.02	35.22
min	0.3	0.3	0.3
max	293.7	202.67	296.3
std	48.03	40.66	41.84
median	33.9	25.35	19.71
Q1	12.3	9.28	5.15
Q3	68.2	58.02	53.22

N: number of NO_3measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

3.4.3 SOM-based specific vulnerability model

The SOM-based vulnerability methodology presented in section 3.3.3.2had also been applied to the regional scale of Catalonia. The clustering capabilities of SOM are used again to explore and identify homogeneous areas and relevant features that best discriminate each vulnerability zone in the physical map without using geographical coordinates.

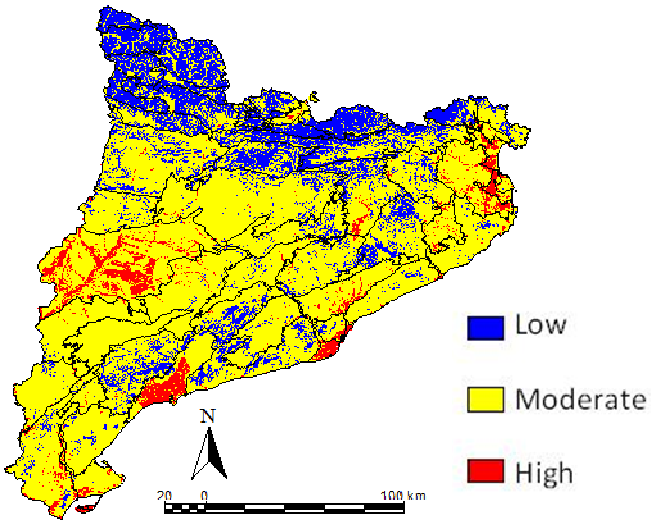


Figure 3.27. SOM-based specific vulnerability map for Catalonia. The variables used to characterize vulnerability are piezometrics, annual rainfall, soil's permeability, surface slopes, aquifer media, hydraulic conductivity, and land uses

A vulnerability map at the regional scale has been generated by self-organizing the seven previously selected hydrogeological and climate properties [piezometric level (H), annual rainfall (P), soil`s permeability (Ks), aquifer media (A), land surface slopes (%s), hydraulic conductivity of the aquifer (Ka) and land uses (land)]. The optimal clustering was obtained using a double-SOM of 4480 and 345 units, with a final quantization and topological error of 0.045 and 0.113, respectively. Figure 3.27 shows the three classes obtained using these variables in the training process of the SOM for Catalonia.

The map obtained is coherent with the cumulative exposure maps shown in Figure 3.25. In addition the SOM based vulnerability index produces vulnerability maps which are compatible with those produced using the DRASTIC index. Nevertheless, the SOM-based vulnerability method is capable of coping with the scarcity of environmental data sets. The SOM provides a consistent framework to infer the missing data and to estimate (interpolate) the exposure concentration of diverse pollutants even from few data. It should be noted, however, that if the amount of information is insufficient the quantization error provides a measure to diagnose the reliability of the SOMinferential process.

Statistics of pollutants concentrations for the vulnerability classes detected in the SOM-based vulnerability map for Catalonia are presented in Tables 3.31 and 3.32. SOM-based vulnerability classes are highly correlated to NO_3 mean concentrations, as the mean values increases with vulnerability, and the "high" vulnerability zone has a mean concentration value above regulatory limit of nitrates. Cumulative point data reveals that "low" vulnerability class includes only 3% of measurement stations exceeding regulatory limits for the area of study.

Table 3.31. Frequency of exceeding cumulative legal threshold at year 2002 and vulnerability class in SOM-based vulnerability map for Catalonia area

#Exd	Low	Mod	High
0	5	173	60
1	12	167	43
2	1	95	33
3	1	62	31
4	0	28	16
5	1	7	7
6	0	4	1

Exd: number of pollutants exceeding legal thresholds in the same monitoring station

Table 3.32.SOM-based specific vulnerability map statistics for NO_3 at year 2002 in Catalonia area

Stats	Low	Mod	High
N	16	462	159
Nexd	3	136	69
mean	33.29	39.31	53.67
min	6.1	0.3	0.3
max	107.1	293.7	296.3
std	27.62	41.81	52.98
median	24.45	25.03	43.07
Q1	14.1	8.5	12.3
Q3	43.5	58.2	77.05

Symbols are the same as in Table 3.30.

Visual comparison of Figures 3.26 (DRASTIC vulnerability) and 3.27 (current*vIndex* approach) indicates that the SOM vulnerability index yields lower vulnerability estimates but with a more consistent spatial continuity than for the DRASTIC model. This is due to the clustering and topology preservation properties of the SOM learning algorithm. The aquifers located at the North-East part of Catalonia are the most vulnerable mainly due to the impact of the high population density areas that are mainly located along the Mediterranean coast. Other vulnerable areas, related to human activities, are located near the most populated cities in Catalonia where most of the Catalan industry and logistic distribution centers are also located.

Figure 3.28(left) presents a zoom of the SOM-based vulnerability map for Catalonia to show only the Camp de Tarragona area previously studied. Comparison of SOM-based vulnerability maps for Camp de Tarragona obtained at the local scale, Figure 3.28(right), and at the regional scale, Figure 3.28(left),shows that there is good agreement after the up-scaling process. The regional scale model provides a first estimate of vulnerable areas to be further evaluated with local scale models by refining data resolution, i.e., by a better zonification of vulnerability classes.

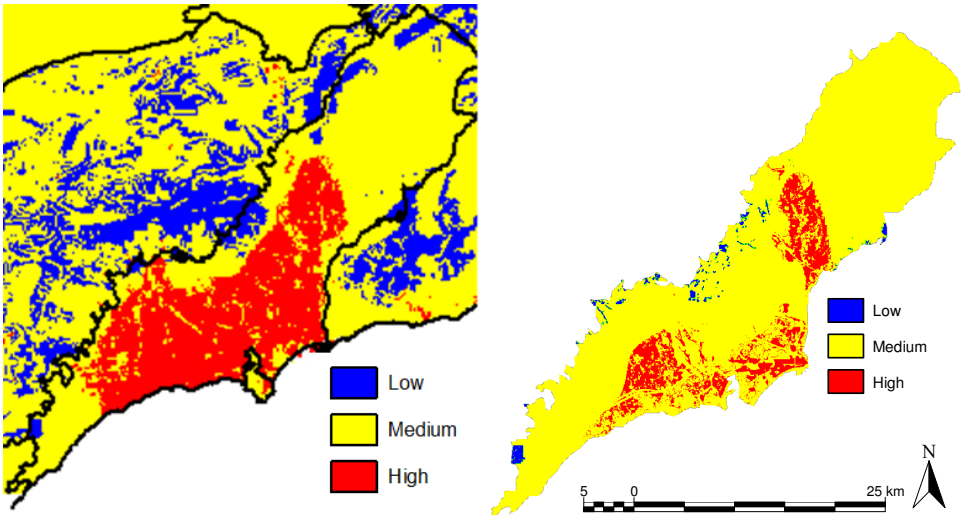


Figure 3.28. (left) Zoom of CataloniaSOM-based specific vulnerability map; (right) Camp de Tarragona SOM-based specific vulnerability map

The proposed SOM method could be easily extended with additional vulnerability indicators to obtain more reliable and detailed vulnerability estimates. The SOM approach also provides a mechanism of knowledge extraction through the use of the C-planes and the distance matrices obtained during the clustering process of hydrogeological and climate parameters for the region considered. By using these complementary information sources, cause and effect relationships between groundwater vulnerability and diverse stressors could be extracted and future contamination in, for example, drinking water supplies minimized. The analysis of these relationships could help in attaining a better understanding of groundwater quality issues and enforcement of adequate urban planning policies in Europe.

3.5 Conclusions

The Self-Organizing Map algorithm provides a consistent framework that can be successfully applied at different levels of the ERA framework and at different spatial resolutions. The proposed methodology for groundwater vulnerability assessment has been tested at two different scales. At the local level in the Camp de Tarragona hydrogeological unit, where the concentration distribution functions of diverse stressors have been determined from available data, the clustering capabilities of the SOM provide vulnerability estimates without requiring previous expert's knowledge ratings of numerical variables. At wider scales such as the regional level of Catalonia, the SOMhas been successfully used to deal with missing data, spatial interpolation, probabilistic risk analysis, and intrinsic and specific vulnerability assessment. This latter aspect is addressed by the development of a new vulnerability index that is able to integrate several sources of diverse hydrogeological and climatic information. This vulnerability index is standardized and discretized in such a way that it is independent of the scale of the geographical region considered. The values obtained from this new index can thus be easily applied to estimate groundwater vulnerability at different scales within Europe. The new methodology provides more detailed and geographically consistent vulnerability maps than the well-established DRASTIC methodology in all its variations proposed previously in the literature.

The SOM approach can help regulators and policy makers to understand the relationships between the potential stressors of concern in an environmental risk scenario such as the pollution of groundwater and the vulnerability of drinking water sources. As the proper management of water resources is becoming a major concern in Europe and in the rest of the world, the proposed SOM based approach for groundwater intrinsic and specific vulnerability assessment provides a reliable and adaptable tool for resource planning and decision making.

Future extensions to this technique should provide mechanisms to automatic tuning the optimal neighborhood used for exposure modeling. Also, the categorization of the vulnerability classes could be exploited to produce vulnerability-driven risk assessment models in which accurate probabilistic risk models could be adjusted to each vulnerability category. Additionally, the use of hierarchical ensembles of SOMs could provide an integrated view of vulnerability at different spatial scales and facilitate the inference of relationships between vulnerability estimates at each scale.

3.6 References

AHMED, A.A. (2009) "Using Generic and Pesticide DRASTIC GIS-based models for vulnerability assessment of the Quaternary aquifer at Sohag, Egypt". Hydrogeology Journal 17: 1203-1217.

ALLER, L., J.H. LEHR, R. PETTY and T. BENNETT (1987) "Drastic: A standardized system to evaluate groundwater pollution potential using hydrogeological settings. United States Environmental Protection Agency. Project Summary EPA/600/S2-87/035.

ALMASRI, M.N. (2008) "Assessment of intrinsic vulnerability to contamination for Gaza coastal aquifer, Palestine". Journal of Environmental Management 88: 577-593.

ANDREO, B.,N. GOLDSCHEIDER, I. VADILLO,J.M. VÍAS, C. NEUKUM, M. SINREICH, P. JIMÉNEZ, J. BRECHENMACHER, F. CARRASCO, H. HÖTZL, M.J. PERLESandF. ZWAHLEN (2005) "Karstgroundwaterprotection: First application of a Pan-European Approach to vulnerability, hazardandriskmapping in the Sierra del Líbar (Southern Spain)". Science of the Total Environment 357: 54-73.

BABIKER, I.S., M.A.A.MOHAMED, T. HIYAMAand K. Kato (2005) "A GIS-basedDRASTIC model for assessingaquifervulnerability in KakamigaharaHeights, GifuPrefecture, central Japan". Science of the Total Environment 345: 127-140.

BURCHART, A., B. LEPPIG,A. MACDONALD, B. MÜLLERand G. WIMMER (2006)"Mappingthegroundwatervulnerability in NorthRhine- Westphalia, Germany". EnvironmentalEngineeringScience 23: 574-578.

CÉRÉGHINO, R. and Y.S. PARK (2009)"Review of the Self-Organizing Map (SOM) approach in water resources: Commentary". Environmental Modeling and Software 24: 945-947.

CHEN, Y., YANG, S., DONG, S., LI, Y., SUN, B., SHAO, Z (2010) "Influence of Agricultural Activity and Aquifer Intrinsic Vulnerability on Groundwater Quality in the Dagu River Watershed (Qingdao, China),"4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE) pp.1-6, 18-20.

DRAOUI, M.,J. VIAS,B. ANDREOand K. TARGUISTI (2008)"A comparative study of four vulnerability mapping methods in a detritic aquifer under Mediterranean conditions". Environmental Geology 54: 455-463.

DIXON, B. (2005a) "Applicability of neuro-fuzzytechniques in predictingground-watervulnerability: A GIS-basedsensitivityanalysis". Journal of Hydrology 309: 17-38.

DIXON, B. (2005b) "Groundwatervulnerabilitymapping: A GISandfuzzyrulebasedintegratedtool". AppliedGeography 25: 327-347.

GEMITZI, A.,C. PETALAS, V.A. TSIHRINTZISand V. PISINARAS (2006)"Assessment of groundwater vulnerability to pollution: a combination of GIS, fuzzylogicanddecisionmakingtechniques". EnvironmentalGeology 49: 653-673.

GOLDSCHEIDER, N. (2005)"Karstgroundwatervulnerabilitymapping: application of a new method in theSwabian Alb, Germany". Journal of Hydrology 13: 1431-2174.

HU, K., Y. HUANG, H. LI, B. LI, D. CHENand R. E. WHITE (2005)"Spatial variability of shallow groundwater level, electrical conductivity and nitrate concentration, and risk assessment of nitrate contamination in North China Plain". Environment International 31: 896-903.

KALTEH, A.M., P. HJORTHand R. BERNDTSSON (2008)"Review of the self-organizing map (SOM) approach in water resources: Analysis, modeling and application". Environmental Modeling and Software 23: 835-845.

KAZEMI, S.M. and S.M.HOSSEINI (2011) "Comparison of spatial interpolation methods for estimating heavy metals in sediments of Caspian Sea". Expert Systems with Applications 38(3): 1632-1649.

KASKI, S. (1997) "Data exploration using self-organizing maps". Department of Computer Science and Technology. Doctor of Technology. Helsinki University of Technology.

KOHONEN, T. (1990) "The self-organizing map". Neurocomputing 21: 1-6.

KOHONEN, T. (2001) "Self-Organizing Maps". Springer-Verlag, Berlin.

LAHR, J. and L. KOOISTRA (2009) "Environmental risk mapping of pollutants: State of the art and communication aspects". Science of the Total Environment 408(18): 3899-3907.

LIGGETT, J.E. and D. ALLEN (2011) "Evaluating the sensitivity of DRASTIC using different data sources, interpretations and mapping approaches". Environmental Earth Science 62: 1577-1595.

LINDSTRÖM, R. (2005) "Groundwater vulnerability assessment using process-based models". Dissertation for the Degree of Doctor in Technology.KungligaTekniskahögskolan.

MACKAY, D., A. DI GUARDO, S. PATERSON, G. KICSI, C.E. COWAN and D. M. KANE (1996) "Assessment of chemical fate in the environment using evaluative, regional and local-scale models: Illustrative application to chlorobenzene and linear alkylbenzenesulfonates". Environmental Toxicology and Chemistry 15(9): 1638-1648.

MAO, Y.-Y., X-G. ZHANGand L. SWANG (2006)"Fuzzypatternrecognitionmethod for assessinggroundwatervulnerability to pollution in theZhangjiarea". Journal of Zhejiang University Science A 7: 1917-1922.

MARTÍNEZ-BASTIDA, J.J., M. ARAUZO and M. VALLADOLID (2010) "Intrinsicandspecificvulnerability of groundwater in central Spain: therisk of nitratepollution". Hydrogeology Journal 18(3): 681-698.

MARTÍNEZ-SANTOS, P., M.R. LLAMAS and P. E. MARTÍNEZ-ALFARO (2008)"Vulnerability assessment of groundwater resources: A modeling-basedapproach to the Mancha Oriental aquifer, Spain". Environmental Modeling & Software 23: 1145-1162.

MASETTI, M., S. STERLACCHINI, C. BALLABIO, A. SORICHETTAand S. POLI (2009)"Influence of threshold value in the use of statistical methods for groundwatervulnerabilityassessment". Science of the Total Environment407: 3836-3846.

MAZARIHIRIART, M., G. CRUZ BELLO, L.A. BOJÓRQUEZ TAPIA, L. JUÁREZMARUSHI, G. ALCANTARLÓPEZ, L.E. MARÍNand E. SOTO GALERA (2003)"Groundwatervulnerabilityassessment for organiccompounds: Fuzzymulticriteriaapproach for Mexico City". Environmental Management 37: 410-421.

MISHIMA, Y. and M. TAKADA (2011) "Evaluation of intrinsic vulnerability to nitrate contamination of groundwater: appropriate fertilizer application management". Environmental Earth Sciences 63: 571-580.

MOORE J. S.(1990) "SEEPAGE: A system for early evaluation of thepollutionpotential of agriculturalgroundwaterenvironments". USDA. SCS, NortheastTechnicalCenter. GeologyTechnicalNote.

NEUKUM, C., H. HÖTZLand T. HIMMELSBACH (2008) "Validation of vulnerabilitymappingmethodsbyfieldinvestigationsandnumerical modeling". HydrogeologyJournal 16: 641-658.

NOLAN, B.T. and K. HITT (2006) "Vulnerability of shallowgroundwateranddrinking-waterwells to nitrate in the United States". EnvironmentalScienceand Technology 40: 7834-7840.

NOLAN, B.T., B.C. RUDDY, K. HITTand D.R. HELSEL (1997) "Risk of nitrate in groundwaters of the United States - A nationalperspective". EnvironmentalScienceand Technology 31: 2229-2236.

PASSUELLO, A., O, CADIACH, Y. PEREZ and SCHUHMACHER, M. (2012) "A spatial multicriteria decision making tool to define the best agricultural areas for sewage amendment". Environment International 38(1): 1:9.

PANAGOPOULOS, G.P., A.K. ANTONAKOSand N.J. LAMBRAKIS (2006) "Optimization of theDRASTICmethod for groundwatervulnerabilityassessment via the use of simple statisticalmethodsandGIS". HydrogeologyJournal 14: 894-911.

PEETERS L., F. BACAO,V. LOBOand A. DASSARGUES (2007)"Exploratory data analysis and clustering of multivariate three-dimensional spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map". Hydrology and Earth System Sciences 11: 1309-1321.

PERLES ROSELLÓ, M.J., J.M. VÍASMARTÍNEZ and B. ANDREO NAVARRO (2009) "Vulnerability of human environment to risk: case of groundwater contamination risk". Environment International 35(2): 325-335:

PISCOPO, G. (2001) "Groundwater vulnerability map explanatory notes". Center of Natural Resources.NWS Department of Land and Water Conservation. Parramatta.

PISTOCCHI, A., J. GROENWOLD, J. LAHR, M. LOOS, M. MUJICA, A.M.J. RAGAS, R. RALLO, S. SALA, U. SCHLINK, K. STREBEL, M. VIGHI, P. VIZCAINO (2011) "Mapping cumulative environmental risks: examples from the EU NoMiracle project". Environmental Modeling& Assessment 16: 119-133.

PONS, X. (2006) "MiraMon Geographic Information System and Remote sensing software". http://www.creaf.uab.es/miramon/

POPESCU, I.C., N. GARDIN, S. BROUYEREandA. DASSARGUES (2008)"Groundwatervulnerabilityassessmentusingphysicallybased modeling: from challenges to pragmaticsolutions". Proceedings of CalibrationandReliability in Groundwater Modeling: Credibility in Modeling, ModelCARE 2007Copenhagen, Denmark 320: 83-88.

RALLO, R. (2007) "Multi-tier framework for the inferential measurement and data-driven modeling". PhD dissertation.UniversitatRovira i Virgili. Spain.

RAHMAN, A. (2008) "A GISbasedDRASTIC model for assessinggroundwatervulnerability in shallowaquifer in Aligarh, India". AppliedGeography 28: 32-53.

SANCHEZ-MARTOS, F.,P.A. AGUILERA, A. GARRIDO-FRENICH,J.A. TORRESand A. PULIDO-BOSCH (2002)"Assessment of groundwater quality by means of self-organizing maps: application in a semiarid area". Environmental Management 30: 716-726.

SECUNDA, S., M.L. COLLIN and A.J. MELLOUL (1998) "Groundwater vulnerability assessment using a composite model combiningDRASTIC with extensive agricultural land use in Israel's Sharon region".Journal of Environmental Management 54(1): 39-57.

SINAN, M. and M.RAZACK (2009) “Anextension to theDRASTIC model to assessgroundwatervulnerability to pollution: application to theHaouzaquifer of Marrakech (Morocco)”. EnvironmentalGeology 57: 349-363.

STIGTER, T.Y., L. RIBEIRO andA.M.M CARVALHO DILL(2006) "Evaluation of an intrinsic and specific vulnerability assessment method in comparison with groundwater Stalinization and nitrate contamination levels in two agricultural regions in the south of Portugal”. Hydrogeology Journal 14: 79-99.

TILAHUN, K. and B.J. MERKEL (2010)"Assessment of groundwater vulnerability to pollution in Dire Dawa, Ethiopia using DRASTIC".EnvironmentalEarth Sciences 59: 1485-1496.

TUTMEZ, B. and Z. HATIPOGLU (2010) “Comparing two data driven interpolation methods for modeling nitrate distribution in aquifer”. Ecological Informatics 5(4): 311-315.

URICCHIO, V.F.,R. GIORDANOand N. LOPEZ (2004)“A fuzzyknowledge-baseddecisionsupportsystem for groundwaterpollutionriskevaluation”. Journal of Environmental Management 73: 189-197.

VESANTO, J.and E. ALHONIEMI (2000) “Clustering of the self-organizing map”. IEEE Transactions Neural Network 11(3): 586-600.

VESANTO, J., J. HIMBERG., E. ALHONIEMI and J. PARHANKANGAS (2000) “SOM Toolbox for Matlab 5”. Report A57, Helsinki University of Technology. ISBN 951-22-4951-0. http://www.cis.hut.fi/somtoolbox.

VÍAS J., B. ANDREO, N. RAUBAR and H. HÖTZL (2010) “Mapping the vulnerability of groundwater to the contamination of four carbonate aquifers in Europe”. Journal of Environmental Management 91(7): 1500-1510.

VILLA, F.and H. MCLEOD (2002)“Environmental Vulnerability Indicators for Environmental Planning and Decision-Making: Guidelines and Applications”. Environmental Management 29: 335-348.

WORRALL, F. andT. BESIEN (2005)“The vulnerability of groundwater to pesticide contamination estimated directly from observations of presence or absence in wells”. Journalof Hydrology 303: 95-107.

WORRALL, F., T. BESIEN and D.W. KOLPIN (2002) "Groundwater vulnerability: interactions of chemical and site properties." Science of the Total Environment 299(1-3): 131-143.

WORRALL, F. and D.W. KOLPIN (2003)"Direct assessment of groundwater vulnerability from single observations of multiple contaminants". Water Resources Research 39: 1345-1345.

ZABEO, A., L. PIZZOL, P. AGOSTINI, A. CRITTO, S. GIOVE and A. MARCOMINI (2011) "Regional risk assessment for contaminates sites Part 1: Vulnerability assessment by multicriteria decision analysis". Environment International 37:1295-1306.

ZAREA. H., V.M. BAYAT and A.P. DANESHKARE (2011) "Forecasting nitrate concentration in groundwater using artificial neural network and linear regression models". International Agrophysics 25: 187-192.

Chapter 4

Lead Exposure Assessment

4.1 Introduction

Lead was classified as category one pollutant in the first community program of action in environmental matters (Generalitat de Catalunya, 1997), so it was selected as prototype pollutant in the development of exposure assessment in groundwater by finding cause-effect relationship with sources of lead pollution and hydrogeological attenuation properties in order to predict lead contamination in groundwater.

Lead is a high density element which is soft, flexible and malleable. It is a poor conductor of electricity but a good insulator of sound, vibrations and radiation. It is highly resistant to corrosion but reacts with nitric acid, and with boiling hydrochloric or sulfuric acids. It is not attacked by pure water although dissolved impurities in water can result in a small amount of corrosion. In the presence of oxygen weak organic acids will attack lead. The United States Environmental Protection Agency (EPA) classifies lead as a probable human carcinogen (EPA, 2004).

Lead has a great number of industrial applications, both in its elemental form and in the form of alloys and compounds. The major use of lead is in the manufacture of lead accumulators (responsible for 68% of consumption and growing in parallel with the increase in the number of automobiles), in which it is used both in the metallic form and as lead oxide. Lead compounds have a number of important uses. Lead oxide is incorporated into glass to prevent the escape of radiation from cathode ray tubes (for example televisions and computer screens) and for the manufacture of crystal glass. Lead glazes are used for hygienic scratch-free surfaces on ceramic products and organic lead salts are added to PVC as stabilizers to protect it against degradation (EPA, 2006). The use of lead in paint has virtually disappeared, with lead carbonate and lead sulfate pigments no longer being permitted in paints in the European Union and restrictions to the use of lead in various types of electronic and electrical equipment(EU, 2003).

Water for potable supplies is normally derived from surface freshwater or groundwater sources. Water treatment prior to distribution does not normally add to the lead content, and usually reduces it (often by as much as 50 %). Plumbing systems may contain lead pipes, lead soldering and bronze or brass fittings. Corrosions of these materials, aggravated by water with low pH, and subsequent leaching into the drinking water can contribute significant quantities of lead in systems where these materials are used.

Lead deposited from the atmosphere can enter aquatic systems through direct fallout or through erosion of soil particles. Infiltration of rainwater into groundwater and entry into aquifers normally involve passage through soil. Rainwater can contain appreciable concentrations of lead. These, however, diminish on passage through the soil, as lead binds to soil minerals and humus. Groundwater therefore normally contains very low concentrations of lead.

Anthropogenic input of lead to aquatic ecosystems can occur from sources such as effluents from mining, smelting, refining and manufacturing processes or the dumping of sewage sludge and atmospheric fallout. In general, there is a little correlation between lead concentrations in rain and snow and concentrations in streams that drain watersheds.

In the European Union, the UE Directive 98/70/CE established a date-line to sell lead-added gasoline in member states on January 1st, 2000. Spain adopted the UE Directive on august 1st, 2001. Up to that year, combustion vehicle emissions were an important source of lead released to the environment in Spain. Even though lead emissions of combustion vehicles are not a pollution source nowadays, lead pollution in groundwater bodies has been affected by the cumulative effect of past emissions.

In this chapter, two artificial neural networks were selected to perform lead exposure assessment in groundwater: back-propagation neural network (Rumelhart et al., 1986) and fuzzy ARTMAP (Carpenter et al., 1992). Fuzzy ARTMAP neural classifier enables multi-variable cluster analysis with cause-effects relationships relevant for exposure and human health (Espinosa et al., 2002).

4.2 Area of study and data

Catalonia region was selected for study due the quality and quantity of available data for lead concentration in groundwater. The Catalonia autonomous community in Spain is formed by 49 hydro-geological units (Figure 4.1). The area analyzed covers 32,114 km^{2}and

includes 946 counties, which form independent administrative units. The area has a very dynamic economy with very important industrial, tourism related and agricultural. It includes many important cities, Barcelona, Girona, Tarragona and Lleida, airports and industrial harbors. The total number of inhabitants is approximately 6,360,000. Detailed description of the area of study is presented in section 3.2 of Chapter 3 of this manuscript. Figure 4.2 shows the areal distribution of principal roads and important industrial areas in the Catalonia region.

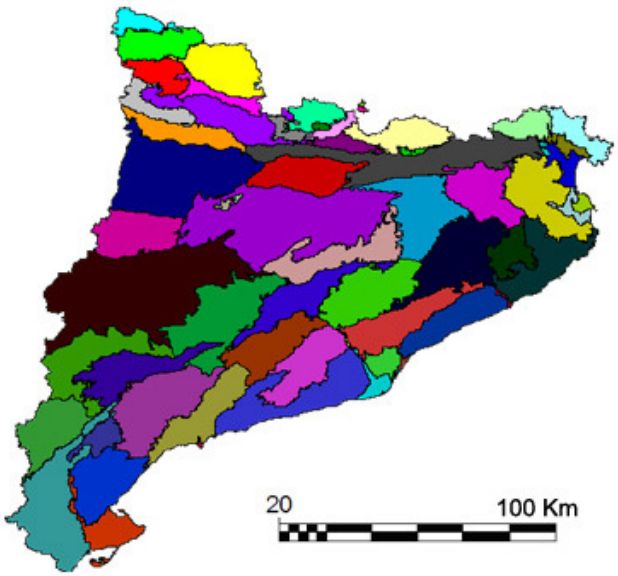


Figure 4.1.Hydrogeological areas in Catalonia region

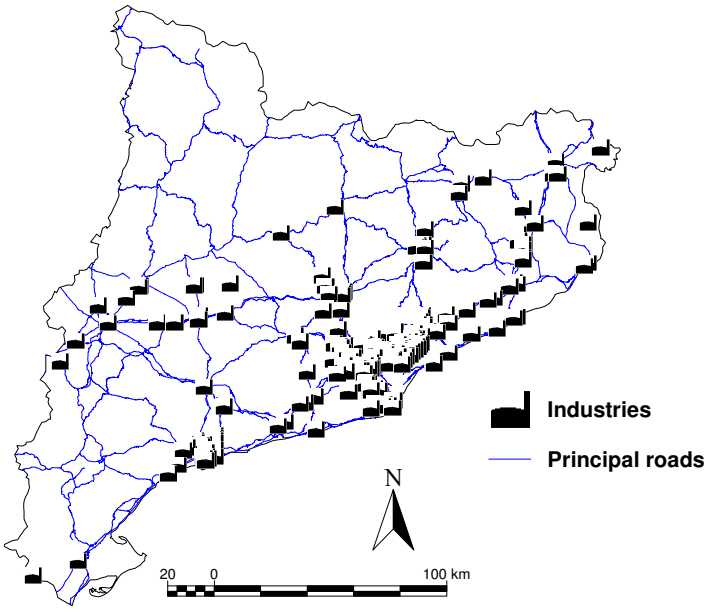


Figure 4.2. Industrial areas and principal roads in Catalonia region for year 2002

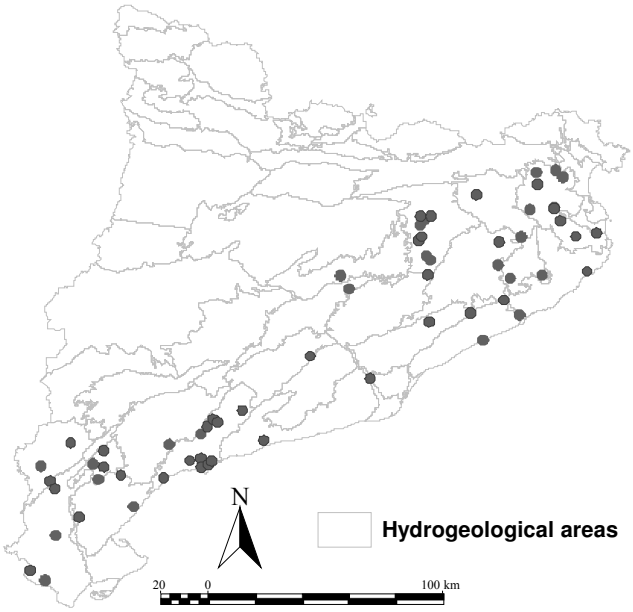


Figure 4.3. Lead measurement points in Catalonia for year 2002 with "detected" value of lead concentration

A total of 401 groundwater lead measures were available from the ACA groundwater quality control network for year 2002. Only 62 measurement points were reported as "detected values" and are located in the eastern side of Catalonia area (Mediterranean side), as is shown in Figure 4.3 (Lead detection threshold in the measurement station is 5µg/l).

4.3 SOM for variable selection

Artificial neural networks like fuzzy ARTMAP (FAM) and backpropagation (BP) needs two set of data to perform the training and test steps. Training data set should have enough information of the global characteristics of the data and cover the whole range.

Self-organizing maps were used to select training and test sets needed to optimize the FAM and BP neural networks. Six variables were identified to perform the cause-effect lead assessment in Catalonia: Hydraulic conductivity, depth to water, topography, distance to principal roads, influence of industries and lead concentrations.

Hydrogeological data (hydraulic conductivity and topography) at each lead measure points presented in Figure 4.3 was extracted from available raster layers developed in Chapter 3 for Catalonia area. Depth to water table values for "lead detected" measure points were obtained from piezometric network of Catalan Water Agency (ACA).

Lead sources were taking in account by distance to principal and secondary roads, and area of influence of principal "lead-emission"industries (Figure 4.2). In order to quantify lead mean travel distance from source, 2 km-radio areas were considered for calculating distance to emission sources.

A toroidal SOM was performed considering relevant variables based on previous results and expert criteria. Data was normalized in order to get SOM's better performance. Figure 4.4 displays U-matrix for trained SOM and C-planes for input variables in the output space. Visual inspection of C-planes reveals strong correlation between industries and lead concentrations, as expected. Others variables are not well-correlated to lead concentrations.

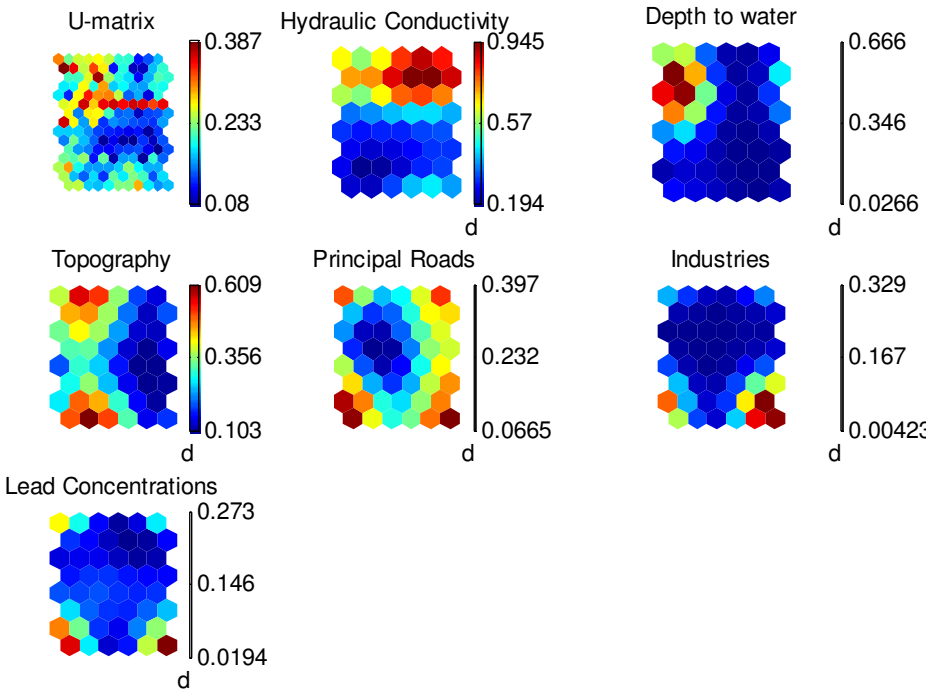


Figure 4.4. U-Matrix (upper -left corner) and distribution in the output space of input variables in the trained SOM

Measure data points were classified in test and training data sets by selecting the centroids of SOM clusters as the training set and the remaining as the test set. Figure 4.5 shows the spatial distribution of training and test sets. Note that due the non-linear classification made by SOM,training points are fewer than test ones. The efficient data feature extraction capabilities of SOM select only relevant prototypes for each cluster (Liu, et al., 2006; Kaski, 1997). In this way, SOM provides an efficient methodology for the selection of optimal training and test sets for the neural networks training and posterior testing as have been demonstrated by Espinosa et al. (2002).

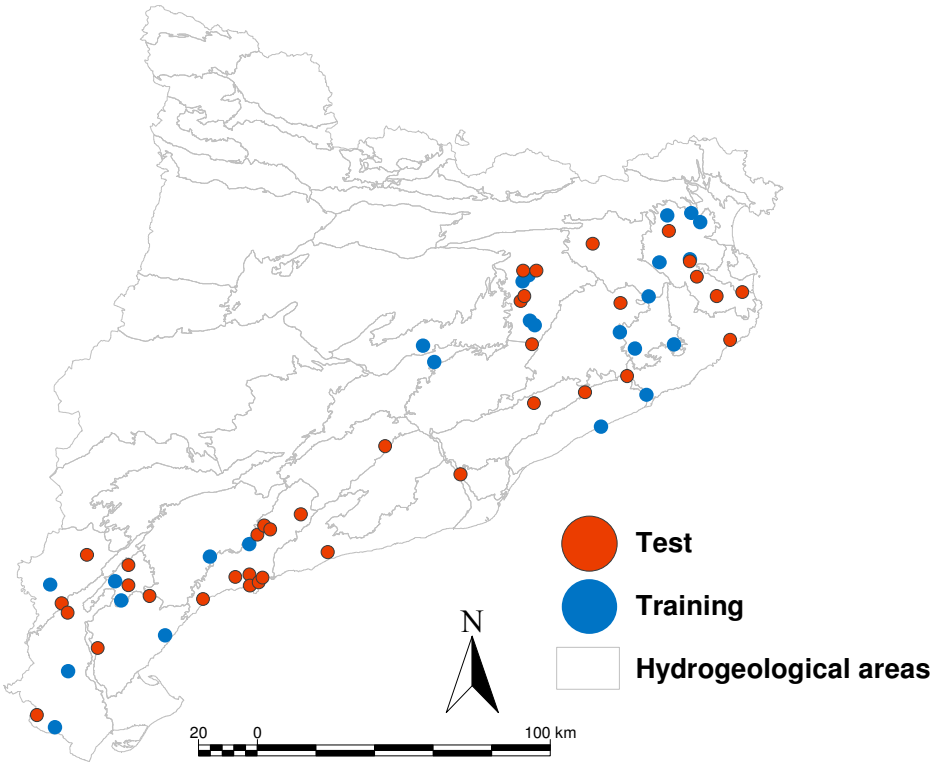


Figure 4.5. Test and training data sets classification by SOM

4.4 Fuzzy ARTMAP for lead exposure assessment

A Fuzzy ARTMAP neural network and a Backpropagation (BP) neural networkwere trained for validate FAM robustness. Network inputs were previously selected by SOM: Hydraulic Conductivity, Depth to water table, Topography, Influence of principal roads and Influence of industrial areas. Output layer is lead measured concentrations in groundwater (Figure 4.6).

Back propagation neural network architecture was optimized by minimizing the arithmetic median error (MAE) for the training set. The final configuration consisted in a three layer feed-forward neural network with 14 neurons in the hidden layer.

Figure 4.7 shows crossplots for trained FAM and BP neural networks for Lead concentration in Catalonia area. Raw values of lead concentrations are not efficient classified in the test setfor both FAM and BP neural networks.

The main application of a FAM classifier is to generate a qualitative cause-effect relationship that helps decision-makers when there is a lack of reliable data. For

visualization purpose, the test data set was expressed in a different way, two categories were identified: moderate and severe, whether sample point is bellow or above legal threshold (10 µg/l) respectively. FAM and BP georeferenced predictions for test data set are displayed in Figure 4.8and Figure 4.9 respectively.

Table 4.1 presents the arithmetic mean error (AME) for FAM and BP, both models behave similarly during training phase but significant differences are encountered in test phase. FAM presented lower errors in predictions than BP and demonstrated the capability of FAM to predict cause-effect relationship between lead concentrations and sources and intrinsic attenuation characteristics of underlying media of groundwater target.

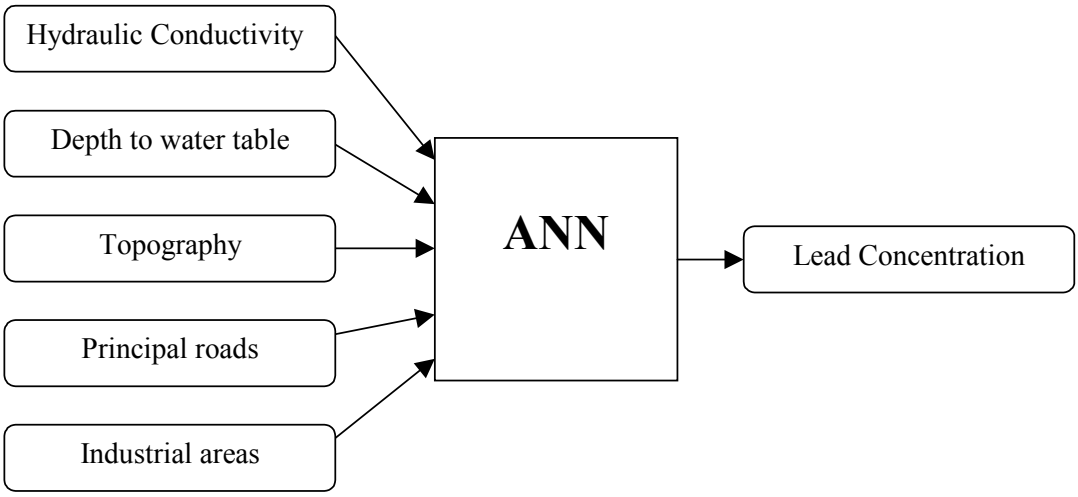


Figure 4.6. Input and output parameters for training a Neural Networks

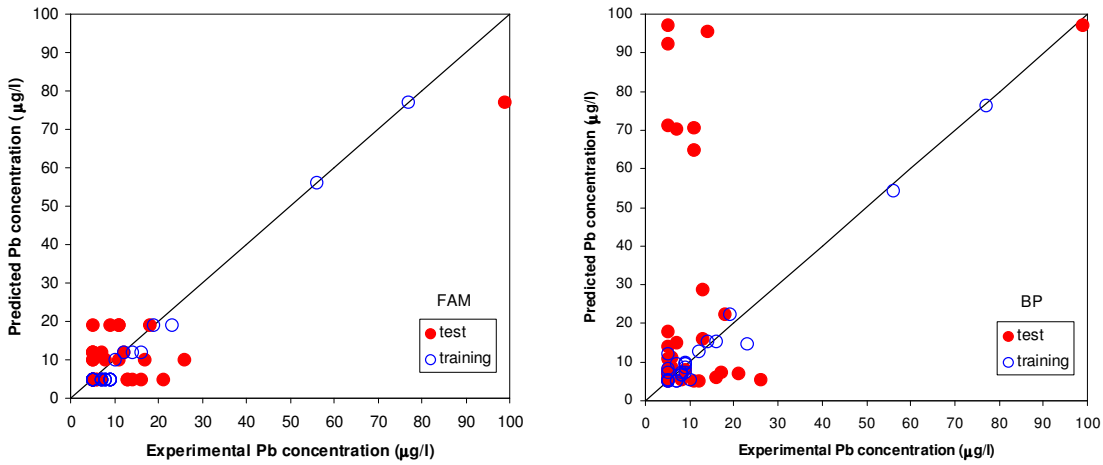


Figure 4.7. Fuzzy ARTMAP (left) and Backpropagation (right) neural networks crossplots for lead concentration (µg/l) in Catalonia area

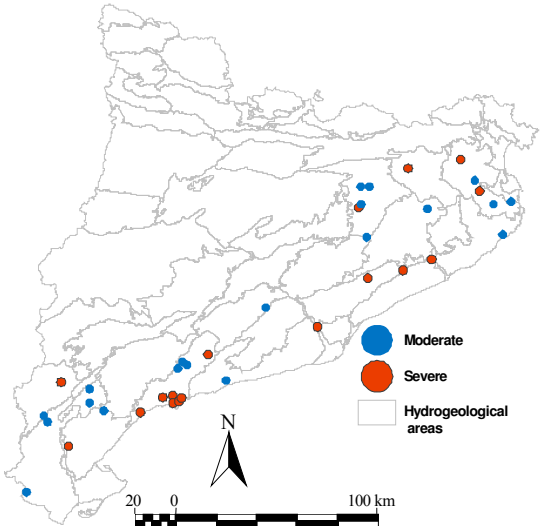


Figure 4.8. Fuzzy ARTMAP test predictions for Lead concentrations in Catalonia area

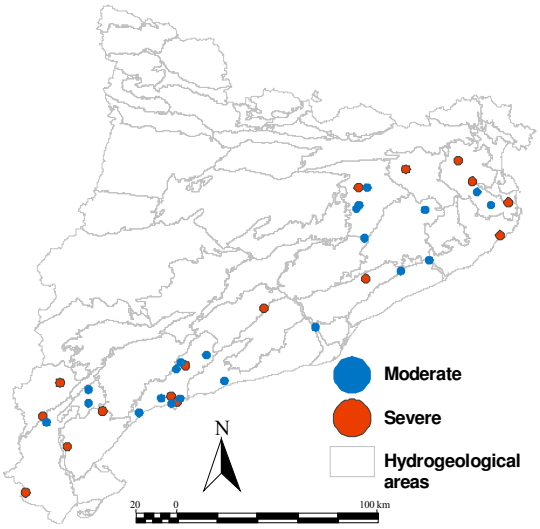


Figure 4.9. Backpropagation network test predictions for Lead concentrations in Catalonia area

Table 4.1. Training and test errors for Fuzzy Art Map and Backpropagation neural networks for Lead concentrations (µg/l) in Catalonia area

	FAM		BP	
Set	AME	std	AME	std
train	1.80	4.89	8.88	6.22
test	1.76	5.65	16.74	6.39

AME: arithmetic median error, std: standard deviation

4.5 Conclusions

Integrated assessment of anthropogenic sources and ecological risks of lead pollution can be achieved by gathering georeferenced information and pollution data and applying this novel methodology using FAM neural classifier.

This part of the study demonstrated that Fuzzy ARTMAP neural classifier is able to establish multi-variable cluster analysis with cause-effects relationships relevant for exposure and human health such as lead exposure assessment in groundwater.

This methodology can be applied to different pollutants in groundwater using proper identification of anthropogenic sources and hydrogeological behavior of the contaminant. Also, cumulative assessment can be accomplished by studying the cause-effect relationships of many pollutants in a specific area or ecological receptor.

Final methodology to assess cause-effect relationships using FAM neural network can be summarized as follows:

(i) Select the target variable or group of variables to study
(ii) Select source/cause variables based on expert knowledge of the process under study
(iii) Normalize source and target variables in a predefined range. Most used ranges are [-1,1] and [0,1].
(iv) Define training and test sets using Self-organizing maps as described in section 4.3.
(v) Train the Fuzzy ARTMAP neural network using the training set and calculate the corresponding error (AME: arithmetic median error).
(vi) Use the test set to evaluate the FAM network performance by calculating the test AME.

Future work can be done to improve FAM performance for lead exposure in Catalonia area by incorporating new variables in the analysis, to account, for example, soil's degradation capacity or climatological variations through the area of study. Spatio-temporal analysis should also be done, by gathering data of different years, and use of SOM as a prediction tool would be evaluated for lead assessment.

4.6 References

BARTFAI, G. (1995) "An improved learning algorithm for the fuzzy ARTMAP neural network". Victoria University of Wellington. Technical Report CS-TR-95/10. Wellington.

CARPENTER, G.A., S. GROSSBERG, N. MARKUZON, H. REYNOLDS and D.B. ROSEN (1992) "Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps." IEEE Transactions on Neural Networks 3: 698-713.

EPA (2004) "Integrated risk information system (IRIS) on lead and compounds (Inorganic)". National Center for Environmental assessment, Office of Research and Development. Washington, DC.

EPA (2006) "Air quality criteria for lead". U. S. Environmental Protection Agency. Final Report EPA/600/R05/144aF-bF. Washington, DC.

ESPINOSA, G., A. ARENAS and F. GIRALT (2002) "An integrated SOM-Fuzzy ARTMAP neural system for the evaluation of toxicity". Journal of Chemical Information and Computer Sciences 42: 343-359.

EU (2003) "Restriction of hazardous substances" European Union Directive 2002/95/EC.

GENERALITAT DE CATALUNYA (1997). "UE/AQ-5.2 EN. Air Quality Daughter Directives. Position Paper on Lead". Environmental Department. General Directorate of Environmental Quality. Commission of the European Communities

KASKI, S. (1997) "Data exploration using self-organizing maps". Department of Computer Science and Technology. Doctor of Technology. Helsinki University of Technology.

KOHONEN, T. (1990) "The self-organizing map." Proceedings IEEE 78(9): 1464-1480.

LIU, Y., R.H. WEISBERG and C.N.K. MOOERS (2006) "Performance evaluation of the self-organizing map for feature extraction". Journal of Geophysical Research, 111: C05018.

OSHA.(2006) "Safety and health topics. Toxic metals. Lead". Occupational Safety and Health Administration, from http://www.osha-slc.gov/SLTC/metalsheavy/index.html.

RUMELHART D.E, G.E. HINTON and R.J. WILLIAMS(1986) "Learning representations by back propagating errors". Nature 323:533-536.

UBILLUS, J. (2003) "Estudio sobre la presencia de plomo en el medio ambiente de Talara en el año 2003". Facultad de Química e Ingeniería Química. To obtain the degree of: Chemical Engineer. Universidad Nacional Mayor de San Marcos.

Chapter 5

Spatio-Temporal Air Quality Assessment

5.1 Introduction

Air pollution over a wide range of spatial scales, from local and regional to global scales, is of growing concern, with ozone, nitrogen oxides and atmospheric particulates typically considered as indicators of air quality. The study of atmospheric pollutants has increased in recent years due to the understanding of their adverse effects on health (W.H.O, 2003) and the implementation of more restrictive environmental regulations. For example, the European directive 1999/30/EU has regulated the progressive decrease of daily and annual limit values for particulate matterin air until 2010 (see Table 5.1).

Table 5.1. Particulate matter (PM) standards in Spain, including total suspended particulate matter (TSP)

Year	2001	2005	2010
Parameter	TSP ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)
Annual limit value	46	40	20
Daily limit value	70	50	50

Particulate air pollution is formed by a mixture of solid and liquid particles suspended in the air. It is convenient to classify particles by their aerodynamic properties because these properties govern their transport and removal from the air as well as their deposition within the respiratory system. Coarse particles haveaerodynamicdiameter greater than 10 micrometers and fine particles,with aerodynamic diameter lower than 10 micrometers (e. g. PM_{10} and $PM_{2.5}$). Comprehensive reports on particulate matter (PM) in Europe have been recently published (Putaud et al., 2004; Van Dingenen et al., 2004). Main important sources of PM_{10} in both urban and rural are: (a) motor vehicles, (b) wood burning stoves and fireplaces, (c) dust from construction, landfills, and agriculture, (d) wildfires and brush/waste burning, (e) industrial sources, (f) windblown dust from open lands (Salvador

et al., 2004; Lenschow et al., 2001; Viana et al., 2008). Sulfate and organic matter are the two main contributors to the annual average PM_{10} and $PM_{2.5}$ mass concentrations. The contribution of nitrates to PM_{10} and $PM_{2.5}$ becomes important when PM_{10}> 50 $\mu g/m^3$. Finally, black carbon contributes 5% – 10% to $PM_{2.5}$ and somewhat less to PM_{10} at all sites, including the natural background sites.

The main sources of anthropogenic particles are located in urban and industrial areas. In urban environments, primary PM is made up of mineral particular matter eroded from the pavement by road traffic and/or resulting from the abrasion of brakes and tires. Industrial activities such as building, mining, and manufacture of cement, ceramics and bricks are typical sources of primary PM. Moreover, coal combustion has also been a traditional source of primary PM (Vardoulakis and Kassomenos, 2008).

A recent analysis and interpretation of PM_{10} levels recorded in regional ambient air quality networks from eastern Spain have shown that the concentration of PM_{10} is strongly influenced by seasonal effects. The main PM events are local urban/industrial pollution events in autumn–winter, and episodes from mid-spring to early-autumn due to Saharan dust contribution to southern and eastern Spain (Rodríguez et al., 2002a; Rodríguez et al., 2002b; Rodríguez et al., 2003; Querol et al., 2001). In urban environments, PM pollution leads to the deterioration of buildings and surfaces caused by the reaction of acidic products with the stone and by soiling of surfaces (Creighton et al., 1990), resulting in black spots in the areas protected from the rain.

Pollutant concentrations in air exhibit a considerable variability in both space and time. Spatio-temporal analysis and mapping is used in a variety of environmental applications. Bayesian Maximum Entropy (BME) provides a realistic representation of the pollutant variation in the spatiotemporal domain, as well as a quantitative assessment of the uncertainty associated to the model (Vyas and Christakos, 1997; Christakos and Serre, 2000; Christakos, 2002; Puangthongthub, et al., 2007; Douaik et al., 2004; Douaik et al., 2005; Savelieva et al., 2005). Air pollution models can be used to detect long-term space and time trends and the location of major air pollution sources (hotspots) within a certain study area. Furthermore, these models are useful for the optimal planning and positioning of pollutant monitoring sites, and to determine if the regulatory environmental standards are met.

BME is a generalization of classical geostatistics methodologies (as kriging) in the frame of modern spatiotemporal geostatistics (Christakos, 2000). BME allows the inclusion of uncertainty in the data (named “soft data”). In this sense, classical kriging is a particular case of BME where only “hard data” (uniquely value at each measure point) is used and the physical knowledge is represented by the covariance function (or the semivariogram function). In order to generate annual concentration maps using classical kriging at several years, the kriging procedure have to be executed for each study year (i.e. the

covariance function has to be calculated and fitted using data for the specific year and interpolation has to be done, this procedure has to be repeated as year under study are needed). In cases where the target is to study the time evolution of a pollutant in several years, the classical kriging approach is a very time consuming task. On the other hand, each map obtained by this methodology is affected only by the number of measurements available at that year. It is possible that in an intermediate time period at a specific area there are no measurements kriging estimates wouldn't be capable to represent the temporal variations in that areas. The use of spatial and temporal variables in the BME covariance model permits a better representation of spatial and temporal variability of data, and this covariance model has to be calculated only one time for the whole period of study, reducing the amount of tasks to be realized.

Self-organizing maps (SOM) have shown their capabilities to study different aspects of data analysis (Kaski et al.,1997; Laine, 2003; Cottrell, 2003; Puangthongthub et al., 2007; Pistocchi et al., 2011). Combination of SOM with temporal data and times series allows to extract some characteristics of data which would be difficult to visualize using others techniques (Cottrell, 2003). BME approach (Christakos et al., 2002) and Self-Organizing Maps (SOM) (Kohonen, 1990) are usedin this work to generate annual concentration maps of PM_{10} (spatio-temporal maps) and to analyze the impact of the quality and amount of data and their spatial and temporal variability. The purpose of this study is to investigate the capabilities of SOM to perform spatio-temporal interpolations of air pollution data.

The effects of PM pollution on human health have been studied over the last decade. These epidemiological studies revealed a relationship between current PM_{10} concentrations in air and the number of premature deaths due to respiratory and cardiovascular diseases (Dockery and Pope, 1996; Christakos et al., 2007; Vardoulakis and Kassomenos, 2008). Exploration of SOM capabilities to extract cause-effect relationship between air pollution and respiratory diseases is presented in Chapter 6 of this thesis.

5.2 Area of study and data

The current study was performed in the Catalonia area (see chapter 3, section 3.2 for extensive information about Catalonia region) with the purpose of generating PM_{10} concentration maps and to evaluate atmospheric pollution over the complete region. Figure 5.1 shows principal PM_{10} anthropogenic pollution sources (except vehicle exhaust that is represented in figure 5.3) over Catalonia, which correspond to the most industrial areas: Barcelona (covering three counties: Barcelonès, Vallès Occidental and Baix

Llobregat) and Tarragona (covering two counties: Tarragonès and Baix Camp). In the Barcelona area, some major sources of PM_{10} pollution are present: 17 out 39 paper/cellulose industries of Catalonia, as indicated in Table 5.1. Table 5.2 points out the demographic distribution which results in the presence of some over populated areas concentrated in small geographic zones. Barcelona metropolitan area concentrates around the 54% of the total population in only 4% of the total territory (data from year 2006).

Table 5.2. Statistics of industrial sources of PM_{10} in Catalonia in 2007

Region	Number gas/distribution stations	Number power stations	Number paper/cellulose	Number pipeline fuel jetty
Catalonia	1145	10	39	10
Barcelona area	327	2	17	5
Tarragona area	56	6	0	5

Table 5.3. Population distribution over Catalonia in 2006 (Source: IDESCAT, Institutd'Estadística de Catalonia)

Region	Number of habitants	% total population	Area km^2	Area total %
Catalonia	7134697	100	32106.54	100
Barcelona area	3830957	54	1214.87	3.78
Tarragona area	395983	6	1016.52	3.17

Figure 5.2 depicts the officially reported emission data from Generalitat de Catalunyaregistered in PRTR-ES (Spanish register of emissions and pollutant sources) (MAAM, 2004). The main pollution "hotspots" are also located around Barcelona and Tarragona even though not all pollutant sources are reported and presented in PRTR-ES's register but only the major contaminant sources.

Figure 5.3 shows the mean daily traffic over the principal road/highways in Catalonia for the period 2003-2007. As can be seen in the figure there is a uniform road-highways network all over Catalonia but also Barcelona area presents the highest traffic density of the country (more than 100,000 vehicles per day at monitoring stations located on the main highways). This is also an evidence of the high PM_{10} source present around Barcelona area due to vehicle exhaust.

Daily emission concentration reports for particulate matter (PM_{10}) over Catalonia were obtained from the Departament de Medi Ambient of Generalitat de Catalonia. PM_{10} concentrations were measured in a total of 86 monitoring stations during the period of study (2003-2007). The number of monitoring stations increased up to 69 active monitoring stations at year 2007 (see table 5.4). Figure 5.4 shows that the distribution of monitoring stations over Catalonia is not uniform. Pollution monitoring stations are mainly deployed in Barcelona and Tarragona, the main urban and industrial zones in Catalonia.

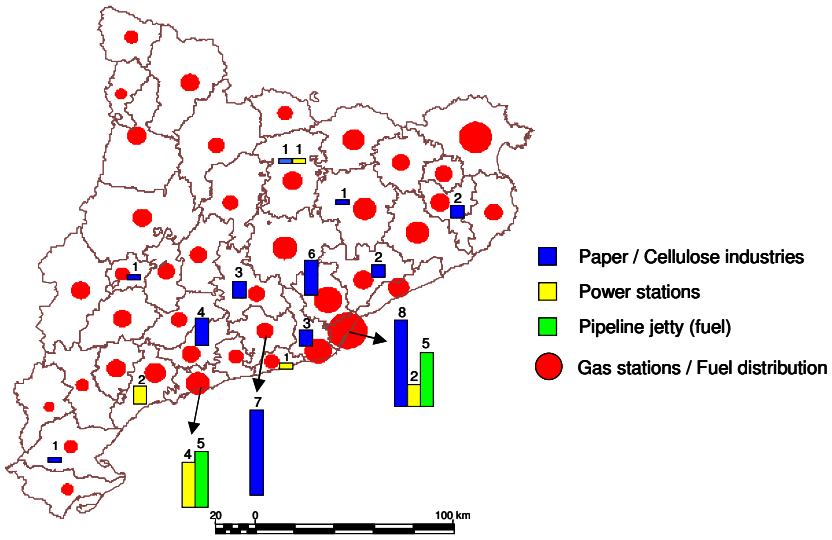


Figure 5.1. Principal anthropogenic pollution sources of PM_{10} in Catalonia at year 2007. The number above each bar indicates the actual number of pollutant sources. Red circles represent pollution sources due to gas/fuel distribution and size is proportional to number of stations

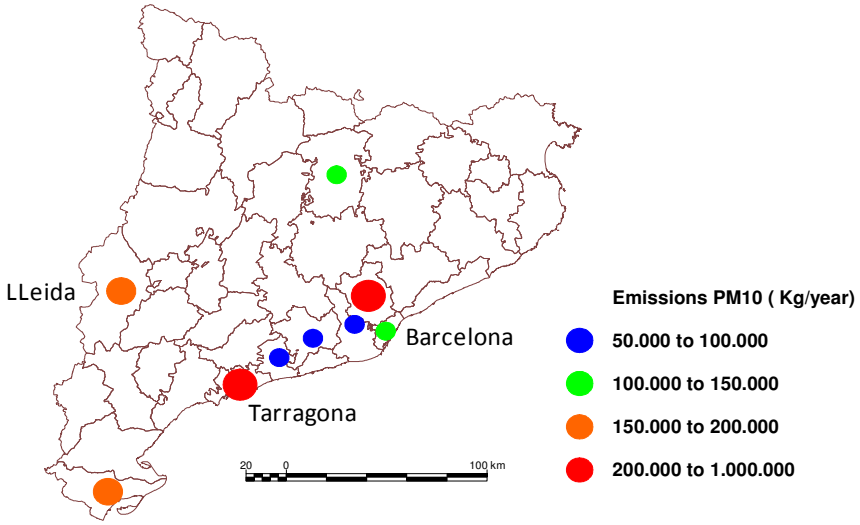


Figure 5.2. PM_{10} emissions for year 2004 reported in the Spanish register of emissions and pollutant sources, PRTR Spain. (Size and color of circles are proportional to level of emissions)

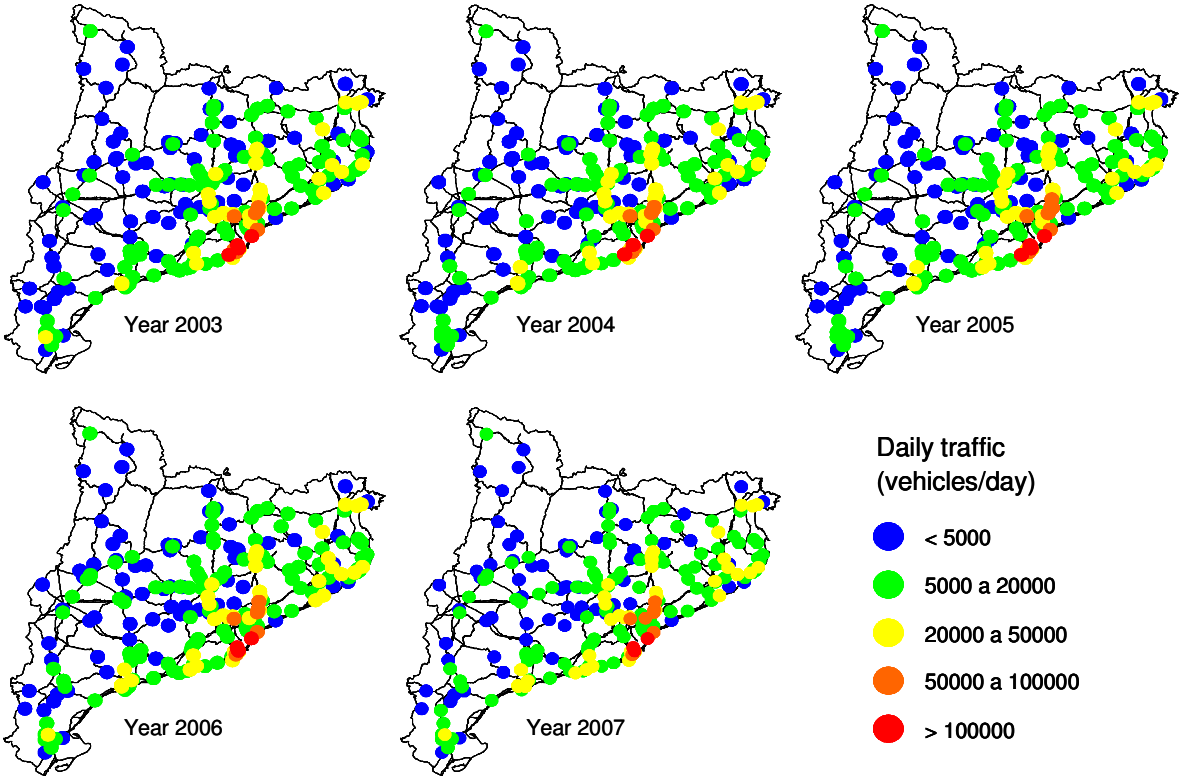


Figure 5.3. Principal roads/highways (black lines) and daily automotive traffic reported at measurements stations in Catalonia for study years (2003 – 2007). (Source: Departament de Política Territorial i ObresPúbliquesfromGeneralitat de Catalonia)

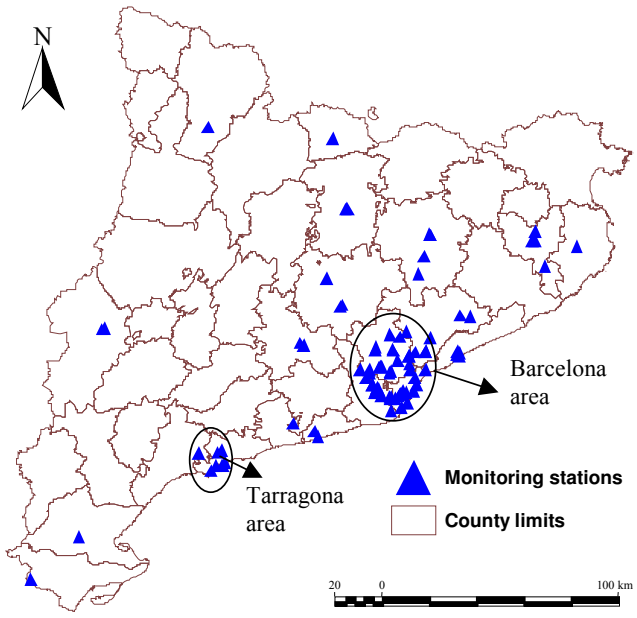


Figure 5.4. PM_{10} monitoring stations in Catalonia over the period of study (2003-2007)

Table 5.4 shows that daily PM_{10} concentrations are incomplete in all monitoring stations over the period studied. The monitoring schedule is not uniform over the calendar year, i.e., monitoring events of stations are not equally time-spaced not even for the same station as is shown in Figure 5.5 and 5.6. The mean number of daily measurements (Ndm) oscillates between 32% and 39% of total annual data (considering a calendar year of 365 days) over the whole period of study. For a monitoring station to be included in the spatio-temporal assessment it should has statistically sufficient data, i.e., a minimum of 80 daily measurements. During 2003 only 37 monitoring stations had Ndm≥80 to contribute to PM_{10} annual mean. This threshold value, Ndm> 80, was only achieved during year 2004for all active monitoring stations (see Figure 5.5). Histograms of daily measurements at each study year presented in Figure 5.6 show that year 2004 is the one with less data dispersion. Also, it is revealed that major frequency of measurements occurred during the first semester of each year of study. Regional seasonal effects cannot be properly predicted using this data.

Table 5.4. Number of monitoring stations (N) and daily PM_{10} measurements (Ndm) in the period 2003-2007. (Percentage of total annual data, based on a 365-days year)

Year	N	min Ndm	mean Ndm	max Ndm
2003	45	7 (2%)	124 (33%)	324 (89%)
2004	37	96 (26%)	142 (39%)	341 (93%)
2005	62	3 (1%)	116 (32%)	289 (79%)
2006	68	48 (13%)	141 (39%)	278 (76%)
2007	69	56 (15%)	121 (33%)	215 (59%)

max: maximum value; min: minimum value; mean: arithmetic mean; std: standard deviation; Ndm: number of daily measurements in a calendar year

Table 5.5. Main statistics of PM_{10} daily average data at study period 2003-2007

Year	max PM_{10} ($\mu g/m^3$)	mean PM_{10} ($\mu g/m^3$)	min PM_{10} ($\mu g/m^3$)	std PM_{10} ($\mu g/m^3$)
2003	187	45.37	3	18.67
2004	400	42.34	8	20.0
2005	232	41.18	2	19.0
2006	389	43.75	2	20.0
2007	473	39.32	2	20.3

max: maximum value; min: minimum value; mean: arithmetic mean; std: standard deviation

Figures 5.7 and 5.8 show the distribution of daily PM_{10} and the logarithm of PM_{10} concentrations at each year, respectively. By applying the log-transform is it possible to generate normally distributed data suitable to be used in spatio-temporal estimations (as can be observed in Figure 5.8). Table 5.5 summarizes the main statistics for daily PM_{10} concentrations. High values of standard deviation shown in Table 5.5 indicate the spatial dispersion of PM_{10} data. Even though, the mean value remains practically constant over the whole period of study (around 40$\mu g/m^3$).

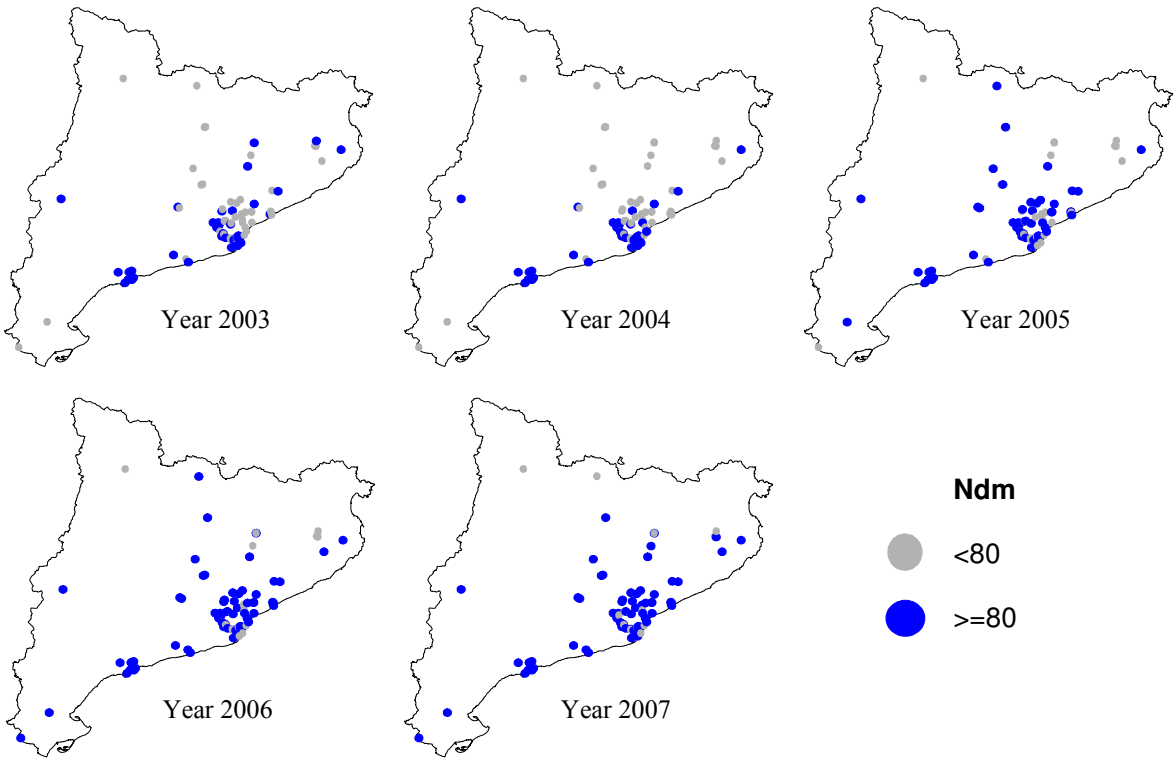


Figure 5.5. Spatial distribution of monitoring stations: based on the number of daily measurements (Ndm). Measurements stations with 80 or more daily measures were considered as statistically sufficient and used in the spatio-temporal interpolation

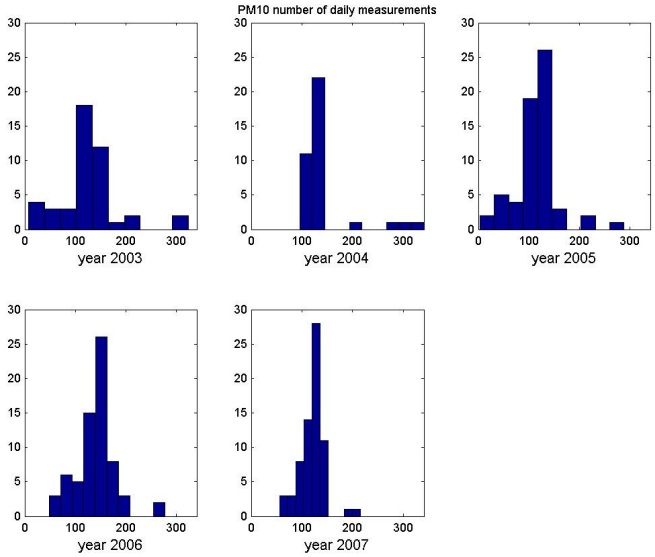


Figure 5.6. Distribution of the number of daily measurements (Ndm) in monitoring stations at each study year (2003-2007)

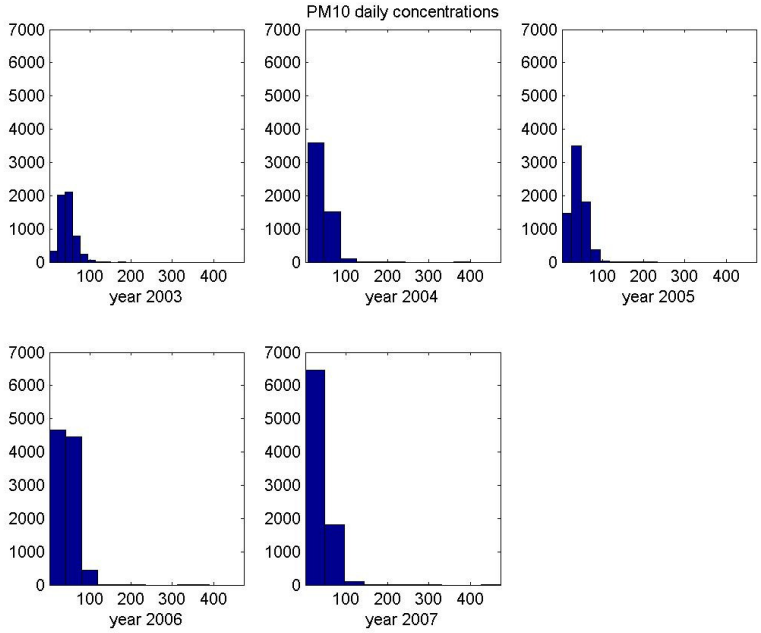


Figure 5.7. Distribution of PM_{10} concentrations in Catalonia at each study year (2003-2007)

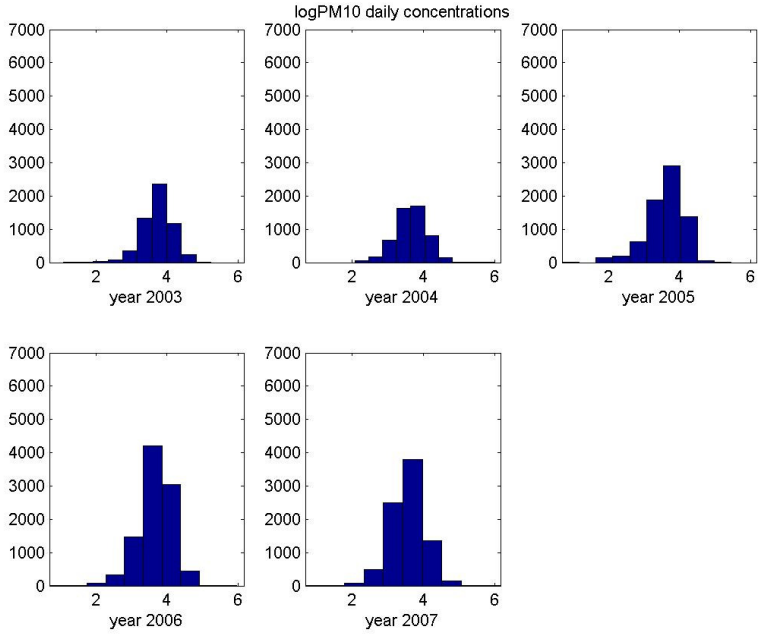


Figure 5.8. Distribution of logarithm of PM_{10} concentrations in Catalonia at each study year (2003-2007)

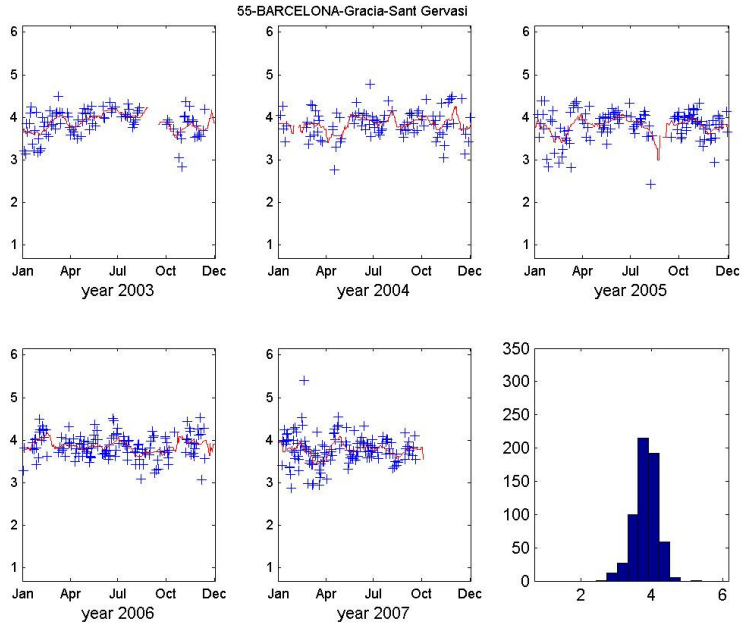


Figure 5.9. LogPM$_{10}$ measurements for years 2003 to 2007 (blue cross), 30-days moving average of logPM$_{10}$ (red line) and histogram of available data for monitoring station 55 (Barcelona- Gracia-SantGervasi)

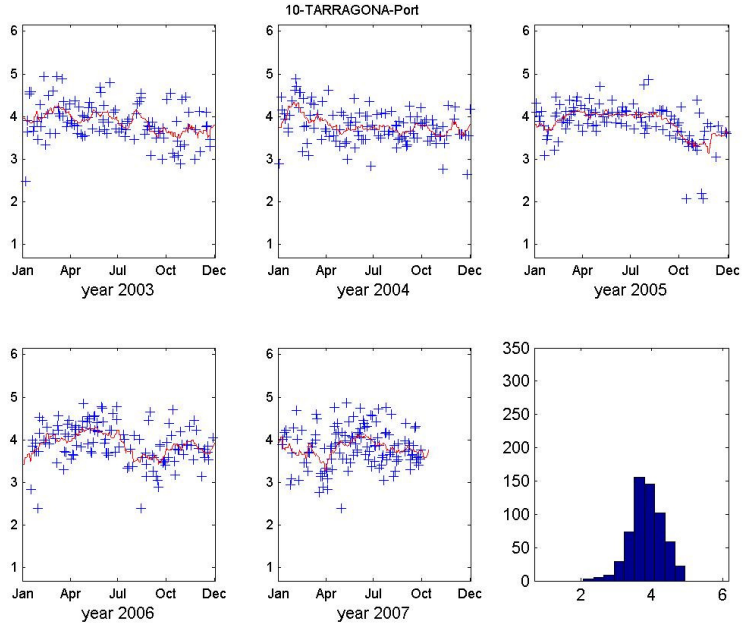


Figure 5.10. LogPM$_{10}$ measurements for years 2003 to 2007 (blue cross), 30-days moving average of logPM$_{10}$ (red line) and histogram of available data for monitoring station 10 (Tarragona - Port)

Annual average values of PM_{10}were computed at each monitoring station using a 30-days moving average approach (based on different window configuration analysis, in those periods where the missing gap is over 30 days no average was calculated remaining the missing gap). Figures 5.9 and 5.10 show the moving average smoothed trends of PM_{10} concentrations for two monitoring stations located in Barcelona area and Tarragona area respectively. Data analysis of the moving averaged trends for the 86 monitoring stations reveals seasonal fluctuations and periodicity at some monitoring station especially at year 2005 but at other time periods there is no conclusive evidence of this periodicity. The evolution of PM_{10} concentration in all monitoring stations is quite similar, having concentration value that oscillates around the regulatory threshold value ($\log[40\mu g/m^3]=3.7$).

Figure 5.11 shows the resulting annual PM_{10} smoothed averages indicating whether or not the regulatory limit (40 $\mu g/m^3$) was exceeded at each monitoring station for each studied year. Exploratory analysis of the data in Figure 5.11 reveals that annual PM_{10} concentrations that exceed regulatory limits are located mainly near the zones that concentrate most of the population (i.e. Barcelona and Tarragona). This is consistent with the stressor effect of population density on PM pollution (mainly due to vehicle exhaust).

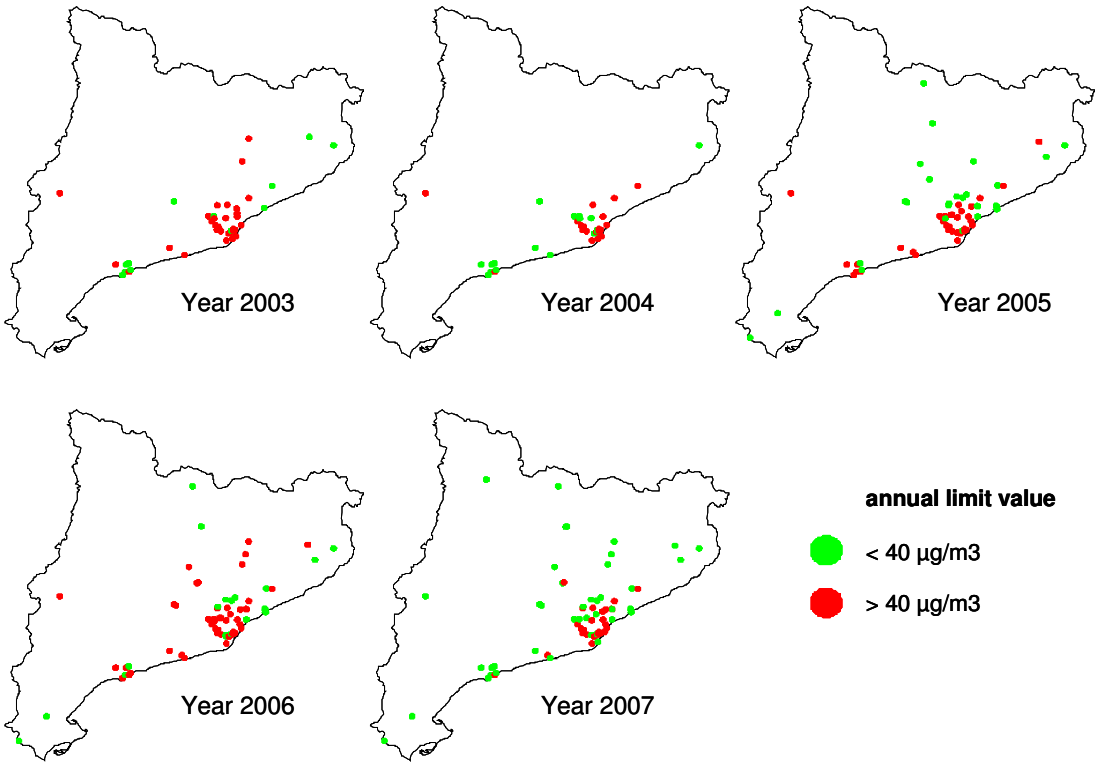


Figure 5.11. Distribution of PM_{10} monitoring stations across Catalonia at each study year (2003-2007) as indicator of exceeding annual limit value (40 $\mu g/m^3$)

5.3 Spatio-temporal mapping

5.3.1 BME approach

Bayesian Maximum Entropy methods have been applied to generate PM_{10} concentration maps for each study year based on PM_{10} data collected at the monitoring stations having more than 80 measurements by year, i.e. blue points shown in Figure 5.5. BME estimators take into account both, space and time using spatial correlations and spatio-temporal covariance function. The spatio temporal (ST) covariance function represents the general knowledge G presented in chapter 2.2.2.

Daily PM_{10} distributions at each monitoring station for each study year are considered as "soft" data in order to generate the corresponding annual PM_{10} concentration estimates. To achieve this goal, an experimental probability density function (PDF) was fitted at each monitoring station for each year of study. Figures 5.12 and 5.13 depict examples of these fitted PDFs for measurements stations 55 and 10 that are located in Barcelona and Tarragona area respectively.

The PM_{10} data set [*Z(s,t)*] was log-transformed [*Y(s,t)=log(Z(s,t)*] and decomposed into a mean trend [$\hat{Y}(s,t)$] and a residual field [*X(s,t)*]:

$$Y(s,t) = \hat{Y}(s,t) + X(s,t) \tag{5.1}$$

The residual field is homogeneous in space and stationary in time having a zero mean. The mean trend is an additive model of spatial and temporal components (Christakos et al., 2001):

$$\hat{Y}(s,t) = m_s(s) + m_t(t) \tag{5.2}$$

where $m_s(s)$ is the spatial component of the mean trend, calculated by applying the moving averaging technique at each monitoring station and $m_t(t)$, the temporal trend, was obtained by averaging the regressed values for each study year.

The theoretical covariance model (c_y) that describes space/time variability of the logarithm of PM_{10} annual mean is a separable model that consists in the superposition of exponential models covering different spatial and temporal scales:

$$c_y(r,\tau) = \sum_{i=1}^{N} c_{0i}\, exp(-3r/a_{si})\, exp(-3\tau/a_{ti}) \qquad (5.3)$$

where r and τ are the spatial and temporal lags respectively, c_{0i} is the spatio-temporal variance, a_{si} and a_{ti} are the spatial and temporal ranges, and N is the number of exponential terms of the covariance.

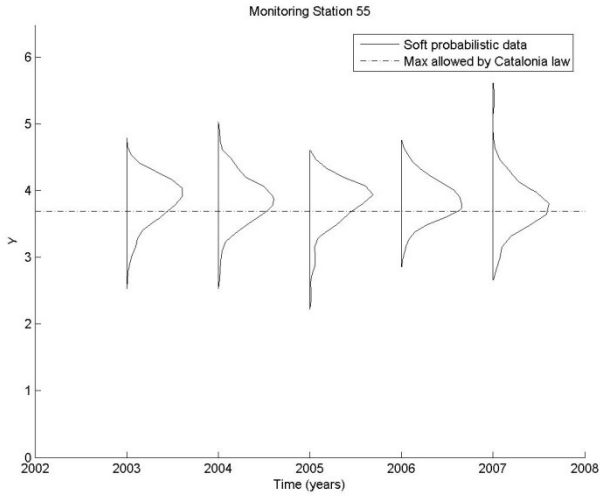


Figure 5.12.Soft probabilistic data of logPM$_{10}$ in monitoring station 55 (Barcelona- Gracia-SantGervasi). Dashed line is the regulatory limit (log[40 μg/m^3]=3.7)

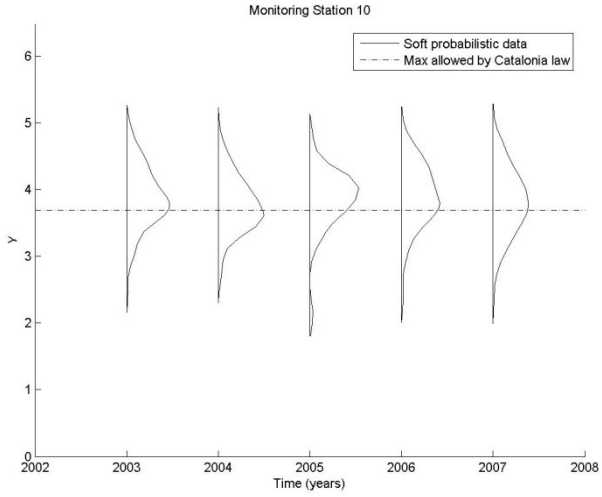


Figure 5.13.Soft probabilistic data of logPM$_{10}$ in monitoring station 10 (Tarragona - Port). Dashed line is the regulatory limit (log[40 μg/m^3]=3.7)

The experimental covariance was calculated using PM_{10} data and several configurations of covariance models were fitted (varying the number of exponential terms). Table 5.6 presents the parameters corresponding to the best covariance configuration obtained using two exponential terms. The first covariance term accounts for the short-range spatial variations (parameter a_{s1} and a_{t1} in table 5.6) and the second term addresses the long-range variations in space and time (parameters a_{s2} and a_{t2} in table 5.6).

Table 5.6. Exponential covariance parameters cases for PM_{10}

c_{01}	a_{s1}	a_{t1}	c_{02}	a_{s2}	a_{t2}
0,0015	5000	1	0,001	10000	2

c_{0i}: spatio-temporal variance; a_{si}: spatial range (meters); a_{ti}: temporal range (years)

To compare the different approaches, a mean standard error (*mse*) was calculated at each monitoring site at each year using the following equation:

$$mse(s,t) = \frac{Z(s,t) - Zest(s,t)}{Z(s,t)} * 100 \qquad (5.4)$$

where Z(s,t) is the measured value at monitoring station located at *s* in time *t* and Zest(*s*,t) is the BME estimation at the same monitoring station. The *total mse* represents the average *mse* for all monitoring stations over the whole period studied (2003-2007).

The BME posterior PDF of annual PM_{10} at each study year over the whole Catalonia area were generated using the Matlab'sBMElib library (Christakos et al., 2002; Bogaert and Serre, 2000). Table 5.7 shows the *mse* for BME interpolation at each study year and the highest *mse* value was obtained at year 2005 where the lowest mean of daily measures occurs (see Table 5.4). The total mse for the whole study period (2003-2007) was 10.84%.

Table 5.7. Main statistics of PM_{10} interpolated data by BME

Year	max PM_{10} ($\mu g/m^3$)	mean PM_{10} ($\mu g/m^3$)	min PM_{10} ($\mu g/m^3$)	std PM_{10} ($\mu g/m^3$)	*mse* (%)
2003	66.49	39.24	18.36	5.56	5.53
2004	61.06	36.02	15.65	5.12	6.19
2005	58.56	34.54	5.60	4.92	16.17
2006	62.86	37.08	5.37	5.29	12.53
2007	56.06	33.06	4.90	4.72	13.8

max: maximum value; mean: arithmetic mean; min: minimum value ; std: standard deviation ; mse: mean standard error

Figure 5.14 shows maps of the BME estimates for the mean annual PM_{10} concentration for the 5 years under study. Visual inspection of BME PM_{10} concentrations reveals that interpolation was better performed inside the imaginary polygons formed by the

monitoring stations where there is continuity in the concentration values. Also, BME is better performed in uniform and dense areas such as Barcelona. Color legend in PM_{10} interpolated maps (for example Figure 5.14) was selected in order to have a visual effect of the differences in concentration values between maps at different years. Resulting polygons are due to the internal algorithm of BME interpolation approach.

Comparison of Figure 5.11 (annual PM_{10} concentration data) and Figure 5.14 (BME estimates) reveals that BME interpolation is able to identify hotspot areas and spatial variability in PM_{10} concentrations among different years. The inclusion of the temporal field in the covariance model affects the prediction of PM_{10} concentrations. Resulting maps reveal the existence of time dependence effects because in areas where there is no monitoring data in a specific year the BME model can predict variability in PM_{10} annual concentration based of prior information of past years.

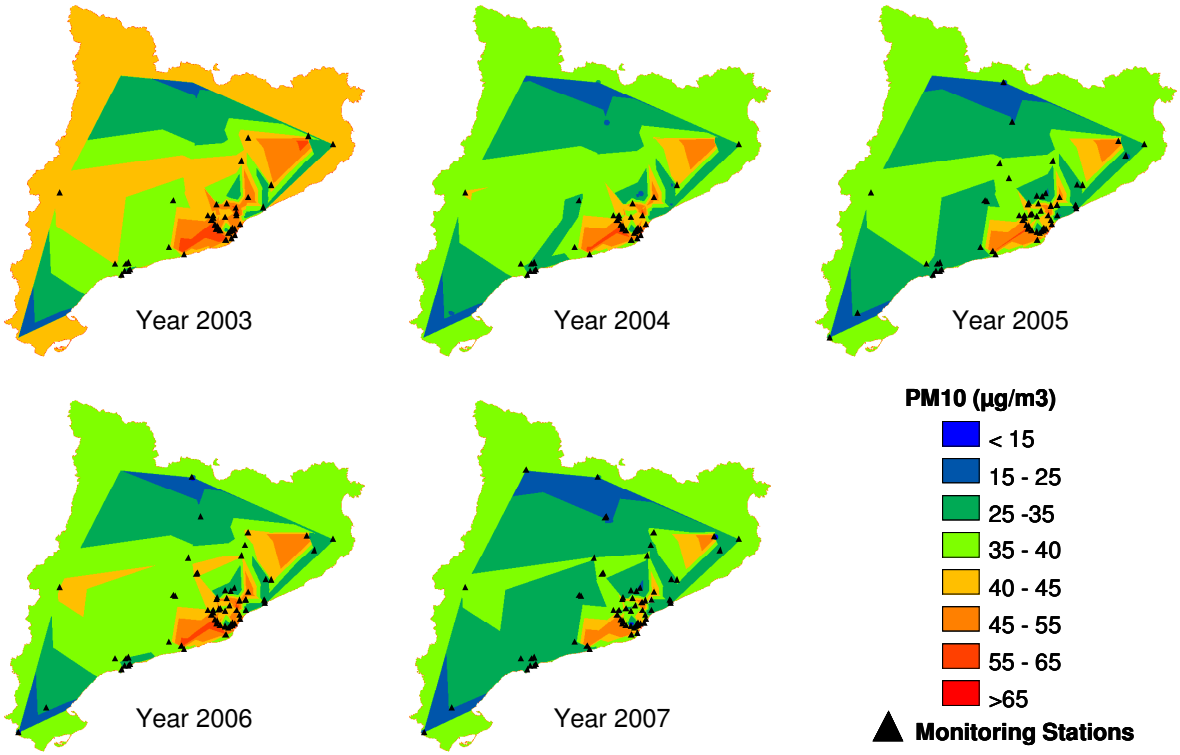


Figure 5.14. BME PM_{10} interpolation over Catalonia at years 2003-2007

The generalization capability of the BME model was investigated by the comparison of BME interpolated values in those monitoring stations excluded in the calculations (those monitoring stations with less than 80 measurements in a year) obtaining a mean error of 8%, 28%, 12% and 17% at years 2003, 2005, 2006 and 2007 respectively (at year 2004 there is no monitoring station having less than 80 measurements). The maximum mean error was obtained at year 2005 where the minimum number of daily measurements is present.

The main limitations of BME models for spatio-temporal mapping are related to the need of defining a general knowledge function (in this case a covariance function) using only target specific knowledge (PM_{10} concentrations measures at monitoring stations). These limitations can be avoided using a different interpolation approach such as the Self-Organizing maps, as presented in next section.

5.3.2 SOM approach

Spatial interpolation using SOM classification capabilities have been demonstrated in chapter 3. In this part of the work, the clustering capabilities of self-organizing maps are used to generate spatio/temporal interpolators of PM_{10} annual concentrations. Different input configurations of SOM inputs were assessed to determine the optimal approach.

Daily PM_{10} data available at each monitoring station were first normalized (in a zero to one scale) and then presented to the SOM to begin the training process. Three main inputs were used: (i) *spatial coordinates* of the monitoring station [UTMX, UTMY], (ii) *year*of the study [2003, 2004, 2005, 2006, 2007] and (iii) *PM_{10} annual mean concentration* at each monitoring station. Temporal component was included in the interpolation by training the SOM with all data available for the time period considered (2003 to 2007) and including explicitly the year in the input vector (T-SOM).

Two approaches of *year* variable representation to training the SOM were studied. First, numerical representation (e.g. YEAR=2003 or YEAR=2006), and second, binary encoded representation (e.g. YEAR_CODE=[1 0 0 0 0] for 2003, or YEAR_CODE= [0 0 0 1 0] for 2006). Also, the inclusion of PM_{10} in the training process was evaluated. In order emulate spatio/temporal kriging calculations, PM_{10} variable was masked for the training process, so it's not considered in the distance calculation between clusters.

Table 5.8 presents T-SOM case studies, input data characteristics and errors obtained after training the map (qe: quantization error, te: topographic error, *mse*: mean standard error at monitoring locations).In cases 1 and 2 the variable PM_{10} is masked for the training process. Case 1 and 2 differ in time representation. Case 1 the variable *year* is presented as row data and in case 2 *year* variable is binary encoded. Cases 3 and 4 include the *year* variable in the training and they differ in time representation. Total *mse* values from table 5.8 reveals that SOM is capable of identify spatio/temporal relationships only using geographical coordinates and time data. The lowest *mse* corresponds to the configuration of case 1.

Toroidal SOMs were selected for spatio-temporal interpolation. Different SOM's size configurations were performed and evaluated using Matlab's SOM Toolbox (Vesantoand Alhoniemi, 2000; Vesanto et al., 2000).

Table 5.8. T-SOM interpolation total error for the different cases

Case	Input data	qe	te	*Total mse* (%)
1	[UTMX UTMY YEAR (PM_{10})]	0,0001	0,040	2,02
2	[UTMX UTMY YEAR_CODE (PM_{10})]	0,0001	0,069	2,20
3	[UTMX UTMY YEAR PM_{10}]	0,010	0,031	23,75
4	[UTMX UTMY YEAR_CODE PM_{10}]	0,010	0,048	26,77

(PM_{10}) masked in training calculation; qe: quantizationerror ;te: topological error; mse: mean standard error

Table 5.9. Main statistics of PM_{10} interpolated data by T-SOM case 1

Year	max PM_{10} ($\mu g/m^3$)	mean PM_{10} ($\mu g/m^3$)	min PM_{10} ($\mu g/m^3$)	std PM_{10} ($\mu g/m^3$)	*mse* (%)
2003	63.88	42.16	32.46	3.84	2.80
2004	55.81	38.36	15.03	6.02	0.81
2005	61.03	38.33	15.03	9.70	2.70
2006	68.41	40.31	22.33	8.03	1.88
2007	83.71	32.97	19.68	6.17	1.92

max: maximum value; mean: arithmetic mean; min: minimum value ; std: standard deviation ; mse: mean standard error

Table 5.9 summarizes statistics of T-SOM PM_{10} interpolations over Catalonia for T-SOM case 1. The highest errors were obtained at years 2003 and 2005, which corresponds to the years with lowest of number of daily measurements and monitoring, stations. The means of the interpolated PM_{10} values are very close to the means of the original input data (table 5.5) even though the standard deviation is quite different for all time periods.

Figure 5.15 shows the cumulative density function for the original input data (blue line) and SOM interpolation (red line) for case 1. The CDF in Figure 5.15 shows that the overall interpolation is better performed when the number of monitoring stations increases (for example year 2007). In contrast, the mean standard error has its lowest value at year 2004 due the low dispersion in data and higher number of daily measurements (i.e., more complete data).

Figure 5.16 shows the T-SOM case 1 PM_{10} maps for years 2003 to 2007. The SOM capabilities for spatio-temporal interpolation have been validated using the experimental data in figure 5.11. Comparison with figure 5.11 (mean annual PM_{10} data) shows that the SOM approach clearly identifies hotspot areas as well as the temporal evolution of the monitoring data.

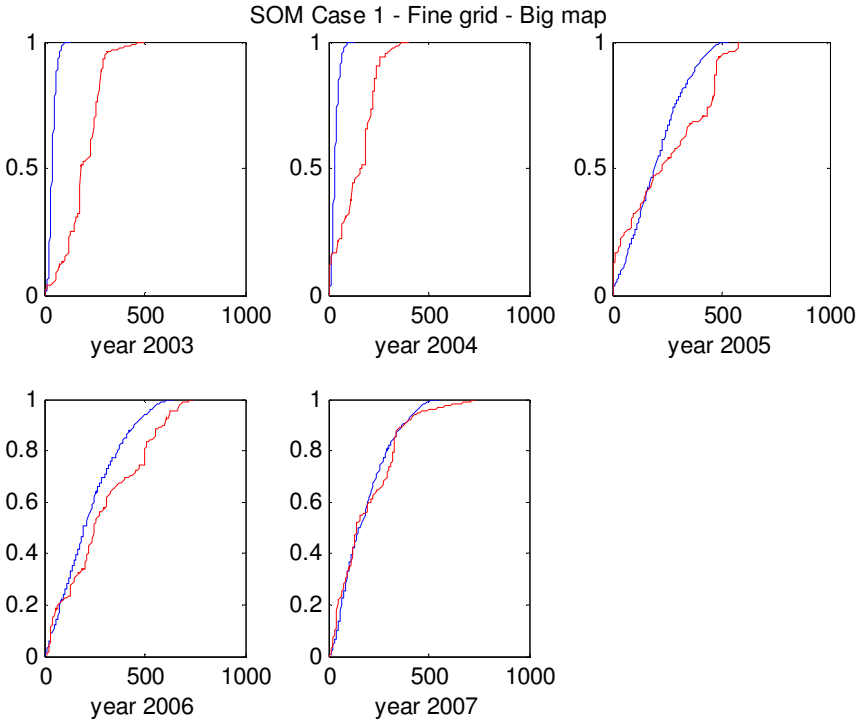


Figure 5.15. Cumulative density function for original data (blue line) and T-SOM case 1 predictions (red line) at each study year (2003-2007)

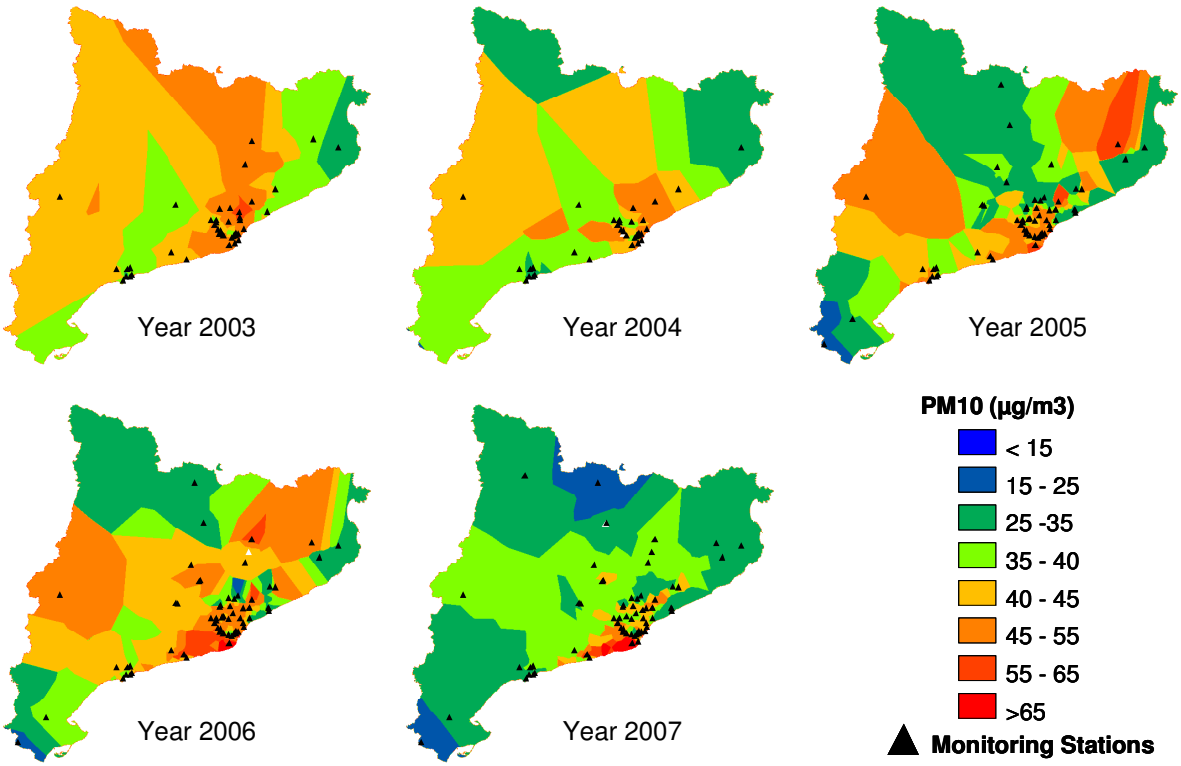


Figure 5.16.PM_{10} interpolation over Catalonia byT-SOM model (Case1)

In order to generate smoother maps, a variation is included in the SOM interpolation process by using a spatial average of 5 best matching units (BMUs) instead the classical method (using only the first BMU) (Rallo et. al., 2005). Table 5.5 presents total *mse* results for different cases using the T-SOM-BMUs approach. Total *mse* errors are larger than those obtained by the classical interpolation method. This is due to the effect of averaging neighbor BMUs to calculate the value in a given spatial point. Figure 5.17 depicts T-SOM-BMUs maps for case 5 at each study year and the corresponding statistics are presented in table 5.10 Concentration maps are similar to those obtained in T-SOM-case 1 (figure 5.16), identifying hot spot areas near Barcelona; using this approach the interpolated map is smoother.

Table 5.10. T-SOM-BMUs interpolation total error for different input vectors

Case	Input data	qe	te	*Total mse* (%)
5	[UTMX UTMY YEAR (PM_{10})]	0.0001	0.040	3.89
6	[UTMX UTMY YEAR_CODE (PM_{10})]	0.0001	0.0069	4.12

qe: quantization error ; te: topological error; mse: mean standard error

Table 5.11.Main statistics of PM_{10} interpolated data by T-SOM-BMUs case 5

Year	max PM_{10} ($\mu g/m^3$)	mean PM_{10} ($\mu g/m^3$)	min PM_{10} ($\mu g/m^3$)	std PM_{10} ($\mu g/m^3$)	*mse* (%)
2003	60.50	42.49	32.43	3.92	4.81
2004	55.80	38.00	25.35	6.06	2.28
2005	60.32	38.14	15.37	9.48	4.78
2006	68.21	40.36	22.32	8.04	3.71
2007	82.37	32.98	19.81	6.19	3.86

max: maximum value; mean: arithmetic mean; min: minimum value ; std: standard deviation ; mse: mean standard error

Statistics of total *mse* are evidence that the T-SOM approach performs better in the identification of spatio-temporal variations in PM_{10} annual concentrations than BME (4% versus 10% of total *mse* for T-SOM and BME respectively). Visual inspection of PM_{10} concentrations maps reveals a better representation of natural variations in the T-SOM approach (figures 5.16 and 5.17) than for the BME (figure 5.14).

Black triangles in Figures 5.16 and 5.17 represent the active monitoring stations for that year. Figures 5.15 and 5.16 confirm the interpolation capability of SOMs despite the missing data. Non-linear clustering capabilities of SOM permit the classification of PM_{10} values taking into account the variability among neighboring monitoring stations and also the intrinsic variability of each monitoring station, extracting by this way relevant information in order to generate annual PM_{10} average concentration.

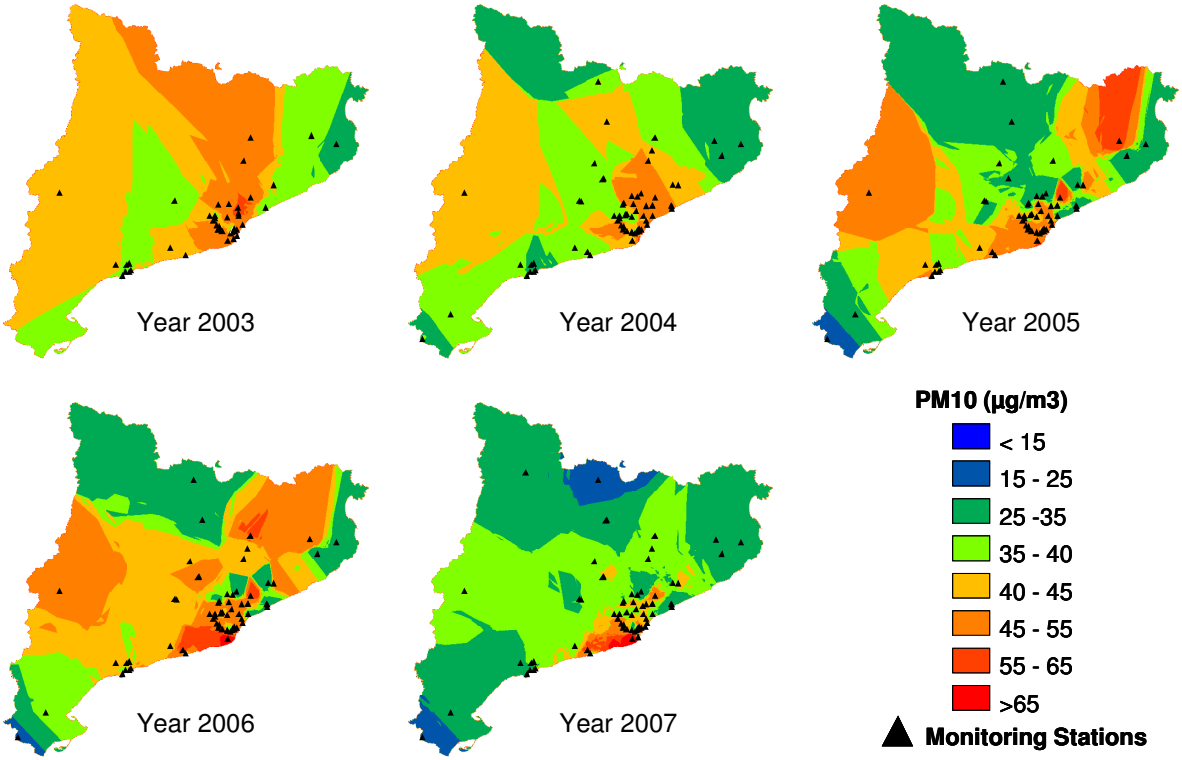


Figure 5.17.PM_{10} interpolation over Catalonia byT-SOM-BMUs (Case 5)

The generalization capability of the SOM model was investigated by the comparison of T-SOM model in those monitoring stations excluded in the calculations (those monitoring stations with less than 80 measurements in a year) obtaining a mean error of 16%, 42%, 31% and 17% at years 2003, 2005, 2006 and 2007 respectively (at year 2004 there is no monitoring station having less than 80 measurements). The maximum mean error was obtained at year 2005 where the minimum number of daily measurements is present. BME generalization test reported (8%, 28%, 12% and 17% at years 2003, 2005, 2006 and 2007 respectively) indicates that the BME approach gives better results than SOM in the case of new data presented to the model (not used in the training set). This is a consequence of having a general knowledge function (the covariance function) in the BME model that it's not present in the SOM approach. Even though, SOM have the benefit that there is no limitation on the minimum number of data points needed in a monitoring station, so all available data can be used in the training phase of the T-SOM model.

5.4 Conclusions

Spatio-temporal maps of PM_{10} annual concentration were developed using two different techniques. Both, thewell-known technique of Bayesian maximum entropy (BME) and the novel technique temporal-Self-organizing map (T-SOM)approaches are capable of identifying hotspot areas in annual PM_{10} concentration maps by interpolating spatio-temporal PM_{10} daily data at monitoring stations over Catalonia.

In the BME approach the use of "soft" data, such as the daily PM_{10} concentration distribution, improves the resulting estimations since it includes temporal data dispersion in the interpolation process (specific knowledge). The need of a general knowledge component, which requires the fitting of a model for the experimental covariance characterization, is a cost-consuming task.

T-SOM has demonstrated to be a very powerful tool to generate annual spatio-temporal maps by extracting relevant information of spatio-temporal daily data (only specific knowledge), without the need of covariance generation.

Future work will be the inclusion of co-variables in the T-SOM training process to improve the accuracy of the spatio-temporal estimations, mainly in unmonitored areas. Some variables to consider might be wind direction, average daily traffic, industrial pollutant sources, and cultivated lands, among others. The final step will be characterization of particulate matter sources using co-variables in T-SOM in order to find non-linear relationships between PM_{10} measurements and different pollution sources.

Exploratory analysis of PM_{10} data in Catalonia evidences important polluted areas around the cities of Barcelona and Tarragona. Also the PM_{10} monitoring network has to be evaluated and redesigned in order to better represent the atmospheric quality status over the whole area. Because of the clustering capabilities of the SOM, it could be used to design a new and optimized monitoring network that take into account the climatological variability and stressors sources through Catalonia area.

5.5 References

BOGAERT P. and M. SERRE (2000) "BMElib for Matlab: The Bayesian Maximum Entropy software for space/time Geostatistics, and temporal GIS data integration". www.unc.edu/depts/BMELIB.

CHRISTAKOS, G. (2000) "Modern spatiotemporal geostatistics". Oxford Univ. Press, NY.

CHRISTAKOS, G.and M. SERRE (2000)"BME analysis of spatiotemporal particulate matter distribution in North Carolina". Atmospheric Environment 34: 3393-3406.

CHRISTAKOS, G., M. SERREandJ. KOVITZ(2001)"BME representation of particulate matter distributions in the state of California on the basis of uncertain measurements".Journal of Geophysical Research106: 9717-9731.

CHRISTAKOS, G., P. BOGAERT andM. SERRE(2002) "Temporal GIS".Springer, New York.

CHRISTAKOS, G. (2002) "On the assimilation of uncertain physical knowledge bases: Bayesian and non-Bayesian techniques". Advances in Water Resources 25: 1257-1274.

CHRISTAKOS, G., R.A. OLEA andH-L.YU (2007)"Recent results on the spatiotemporal modelling and comparative analysis of Black Death and bubonic plague epidemics". Public Health 121: 700-720.

COTTRELL M. (2003) "Some Other Applications of the SOM algorithm: how to use the Kohonen algorithm for forecasting". Invited conference to IWANN 7TH International Work Conference on Artificial and Natural Neural Networks, Menorca, Spain.

CREIGHTON, P.J., P.J. LIOY, F.H. HAYNIE, T.J. LEMMONS, J.L. MILLERandJ. GERHART (1990)"Soiling by atmospheric aerosols in urban industrial area". Journal of the Air & Water Management Association 40: 1285-1289.

DOCKERY, D.andA. POPE(1996)"Epidemiology of acute health effects: Summary of time-series studies". In: Particles in our air: concentration and health effects. Wilson R, Spengler JD (eds), Cambridge, MA. USA, Harvard University Press, pp. 123-147.

DOUAIK, A., M. VAN MEIRVENNE and T. TÓTH (2005)"Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data".Geoderma 128: 234-248.

DOUAIK, A., M. VAN MEIRVENNE,T. TÓTHand M. SERRE (2004) "Space-time mapping of soil salinity using probabilistic bayesian maximum entropy".Stochastic Environmental Research and Risk Assessment 18: 219-227.

KASKI, S. (1997)"Data exploration using self-organizing maps". Department of Computer Science and Technology.Doctor of Technology.Helsinki University of Technology.

KOHONEN, T. (1990) "The self-organizing map". Neurocomputing 21: 1-6.

LAINE, S. (2003) "Using visualization, variable selection and feature extraction to learn from industrial data". Department of Computer Science and Technology.Dissertation for the degree of Doctor of Technology.Helsinki University of Technology.

LENSCHOW, P., H.-J. ABRAHAM, K. KUTZNER, M. LUTZ, J.-D. PREUßand W. RIECHENBÄCHER(2001) "Some ideas about the sources of PM_{10}".Atmospheric Environment 35(S1): S23-S33.

MAAM (2004) "Spanish register of emissions and pollutant sources". www.prtr-es.es.

PISTOCCHI, A., J. GROENWOLD, J. LAHR, M. LOOS, M. MUJICA, A.M.J. RAGAS, R. RALLO, S. SALA, U. SCHLINK, K. STREBEL, M. VIGHI, P. VIZCAINO (2011) "Mapping cumulative environmental risks: examples from the EU NoMiracle project". Environmental Modelling& Assessment 16: 119-133.

PUANGTHONGTHUB, S., S. WANGWONGWATANA, R. KAMENS and M. SERRE(2007) "Modeling the space/time distribution of particulate matter in Thailand and optimizing its monitoring network". Atmospheric Environment 41: 7788-7805.

PUTAUD, J., F. RAES, R. VAN DINGENEN, E. BRUÜGGEMANN, M.C. FACCHINI, S. DECESARI,S. FUZZI, R. GEHRING, C. HÜGLIN, P. LAJ, G. LORBEER, W. MAENHAUT, N. MIHALOPOULOS, K. MÜLLER, X. QUEROL, S. RODRIGUEZ, J. SCHNEIDER, G. SPINDLER, H. BRINK, K. TORSETH and A. WIEDENSOHLER (2004) "A European Aerosol Phenomenology-2: chemical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe». Atmospheric Environment 38: 2579-2595.

QUEROL, X., A. ALASTUEY, S. RODRIGUEZ, F. PLANA, C.R. RUIZ, N. COTS, G. MASSAGUÉandO. PUIG(2001) "PM_{10} and $PM_{2.5}$ source apportionment in the Barcelona Metropolitan area, Catalonia, Spain". Atmospheric Environment 35: 6407-6419.

RODRÍGUEZ, S., X. QUEROL, A. ALASTUEY and E. MANTILLA (2002a) "Origin of high summer PM_{10} and TSP concentrations at rural sites in Eastern Spain". Atmospheric Environment 36: 3101-3112.

RODRÍGUEZ S., X. QUEROL, A. ALASTUEYand F. PLANA (2002b)"Sources and processes affecting levels and composition of atmospheric aerosol in the Western Mediterranean".Journal of Geophysical Research 107(D24): 12-14.

RODRÍGUEZ, S., X. QUEROL, A. ALASTUEY, M.M. VIANA and E. MANTILLA(2003) "Events affecting levels and seasonal evolution of airborne particulate matter concentrations in the Western Mediterranean". Environmental Science and Technology 37(2): 216–222.

SALVADOR, P., B. ATTÍNANO, D.G. ALONSO, X. QUEROL and A. ALASTUEY (2004) "Identification and characterization of sources of PM_{10} in Madrid (Spain) by statistical methods".Atmospheric Environment 38: 435-447.

SAVELIEVA, E., V. DEMYANOV, M. KANEVSKI, M. SERRE and G. CHRISTAKOS (2005)"BME-based uncertainty assessment of the Chernobyl fallout".Geoderma 128: 312-324.

VAN DINGENEN, R., F. RAES, J. PUTAUD, U. BALTENSPERGER, A. CHARRON, M.C. FACCHINI, S. DECESARI, S. FUZZI, R. GEHRING, H.C. HANSSON, R.M. HARRISON, C. HÜGLIN, A.M. JONES, P. LAJ, G. LORBEER, W. MAENHAUT, F. PALMGREN, X. QUEROL, S. RODRIGUEZ, J. SCHNEIDER, H. BRINK, P. TUNVED, K. TORSETH, B. WEHNER, E. WEINGARTNER, A. WIEDENSOHLER and P. WAHLIN (2004) "A European Aerosol Phenomenology-1: physical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe".Atmospheric Environment 38: 2561-2577.

VARDOULAKIS, S. and P. KASSOMENOS (2008) "Sources and factors affecting PM_{10} levels in two European cities: Implications for local air quality management". Atmospheric Environment 42(17): 3949-3963.

VESANTO, J. and E. ALHONIEMI(2000) "Clustering of the self-organizing map".IEEE Transactions Neural Network 11: 586-600.

VESANTO, J., J. HIMBERG., E. ALHONIEMI and J. PARHANKANGAS (2000) "SOM Toolbox for Matlab 5". ISBN 951-22-4951-0. http://www.cis.hut.fi/somtoolbox.

VIANA, M., T. A. J. KUHLBUSCH, X. QUEROL, A. ALASTUEY, R.M. HARRISON, P.K. HOPKE, W. WINIWARTER, M. VALLIUS, M. SZIDAT, A.S.H. PRÉVÔT, C. HUEGLIN, H. BLOEMEN, P. WAHLIN, R. VECCHI, A.I. MIRANDA, A. KASPER-GIEBL, W. MAENHAUT and R. HITZENBERGER (2008)"Source apportionment of particulate matter in Europe: A review of methods and results". Journal of Aerosol Science 39: 827-849.

VYAS V. and G. CHRISTAKOS(1997)"Spatio-temporal analysis and mapping of sulfate deposition data".Atmospheric Environment31(21): 3623-3633.

W.H.O.(2003)"Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide". World Health Organization, Regional Office for Europe,pp.98.

Chapter 6

Human Health Risk Assessment

6.1 Introduction

Environmental risk assessment (ERA) covers the risk to ecosystems (ecological risk assessment) like air, water, land and biological species, and risk to humans, exposed or impacted as defined by European Environment Agency (EEA, 1998). Human health risk assessment (HHRA) is defined as a procedure that defines the probability and magnitude of adverse effects to human health posed by environmental agents.

Epidemiology is the science that studies the distributions and patterns of health-related events (disease, injury, etc.) and their causes in a defined human population. The effects of PM pollution on human health have been studied over the last decade (Kampa and Castanas, 2008). These epidemiological studies revealed a relationship between current PM_{10} concentrations in air and the number of premature deaths due to respiratory and cardiovascular diseases (Dockery and Pope, 1996; Christakos et al., 2007; Vardoulakis and Kassomenos, 2008). Also, studies on relationship in hospital admissions by cardiovascular and respiratory diseases and fine particles air pollution have been developed in recent years (Schwartz and Morris, 1995; Dominici et al., 2006; Chang et al., 2007; Linping et al., 2007; Bell et al., 2009). Some studies had revealed that health effects of air pollution are exacerbated by the coincidence of multiple pollutants exposure (Dominici et al., 2010).

Many epidemiological studies have been focused specially on cause-effect relationship between air pollution and asthma in children, and there is a general consensus on the exacerbation of asthma and allergies in children due to air pollution exposure (Jerrett et al., 2008; Delfino et al., 2008; Delfino et al., 2009; Clark et al., 2010; Gehring et al., 2010; Samoli et al., 2011). Delfino et al. (2009) estimated the association of local traffic–generated air pollution (nitrogen oxides and carbon monoxide) with repeated hospital encounters for asthma in children and concluded that locally generated air pollution near the home affects asthma severity in children.

The present study is focused in the exploration of self-organizing maps (SOM) capabilities to extract cause-effect relationships between air pollution and human health effects. The study was developed in Catalonia area and based on air pollution data and emergency hospital admissions data for years 2004 and 2005. Air quality standards in Catalonia are regulated by Spanish Real Decreto 1073/2002 and 1796/2003. Regulatory values for health-related air pollutants are presented in Table 6.1.

Table 6.1. Air pollutants values in Spain for the protection of human health

Pollutant	units	1-h	8-h	24-h	annual
PM_{10}	$\mu g/m^3$	-	-	50(a)	40
O_3	$\mu g/m^3$	-	120(b)	-	-
SO_2	$\mu g/m^3$	350(c)	-	125(d)	-
NO_2	$\mu g/m^3$	200(e)		-	40
C_6H_6	$\mu g/m^3$	-	-	-	5
CO	mg/m^3	-	10(f)	-	-

(a) It can't be exceeded more than 35 times in a year
(b) Maximum value of 8-h mean during a day. It can't be exceeded more than 25 times in three-year period
(c) It can't be exceeded more than 24 times in a year
(d) It can't be exceeded more than 3 times in a year
(e) It can't be exceeded more than 18 times in a year
(f) Maximum value of 8-h mean during a day

The aim of this study was not an epidemiology study that attempted to demonstrated negative effects in human health by air pollution, because it's assumed that this relationship has been validated elsewhere (W.H.O, 2003). The focus of this work is to evaluate the feature extraction capabilities of self-organizing maps to generate human health risk maps based on air pollution data. Asthma in children (up to 14 years old) was selected to perform HHRA using self-organizing maps in Catalonia area due the quality and quantity of available data.

6.2 Area of study and data

The human health risk assessment of air pollution using self-organizing maps was developed for Catalonia area (detailed description of Catalonia is presented in Chapter 3 and 5 of this manuscript). Demography data for Catalonia was obtained from IDESCAT (Catalan statistics institute). Age-segregated population data is presented in Table 6.2. Selection of age-aggregation ranges was according available health-related data. Catalonia is administrative divided in counties and municipalities. There are 41 counties and 947 municipalities in Catalonia (see Figure 6.1). Figure 6.2 presents areal distribution

of numbers of children (0-14 years old) for municipalities in Catalonia for the study period (2004-2005). Municipal division permits detailed segregated data for spatial analysis.

Table 6.2. Number of habitants by age range in Catalonia at years 2004 and 2005

year	Age range		
	0-14	15-65	>65
2004	950482	4710344	1152493
2005	988016	4856466	1150724

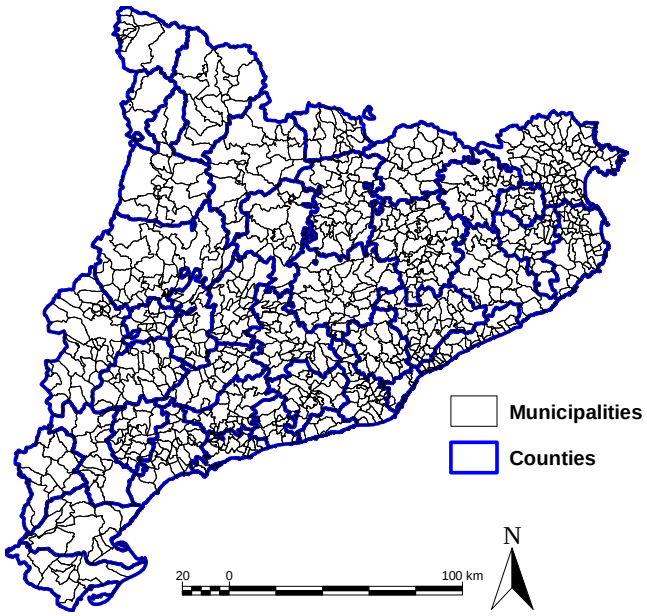


Figure 6.1. Territorial division of Catalonia (blue) county division (black) municipal division

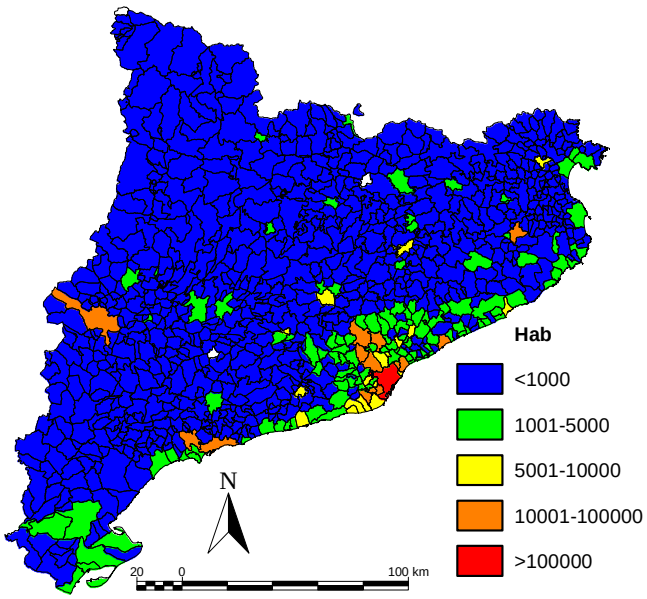


Figure 6.2. Distribution of the number of 0-14 years old habitants in Catalonia area by municipal division (years 2004 and 2005)

Population health-related information data was obtained from CatSalud (Catalan health department). Available information of hospital admissions by respiratory diseases and heart attacks during years 2004 and 2005 was used in this study. Health data was segregated by residence municipality, age range and sex using ICD-9 (International statistical classification of diseases and related health problems).

Summary of health-related available data during the study is presented in Tables 6.3 to 6.5. Respiratory disease data includes: chronic obstructive pulmonary diseases (COPD) and allied conditions (ICD-9 codes 490-496) and pneumoconiosis and other lung diseases due to external agents (ICD-9 codes 500-508). Asthma-related data is classified as extrinsic, intrinsic and chronic obstructive asthma (ICD-9 codes 493.0, 493.1 and 493.2 respectively). Heart disease data consisted in acute myocardial infarction (ICD-9 code 410) and other acute and sub-acute forms of ischemic heart disease (ICD-9 code 411).

Table 6.3. Distribution of number of hospitalization by respiratory diseases by age range in Catalonia at years 2004 and 2005

year	Age range		
	0-14	15-65	>65
2004	1233	8089	24613
2005	1061	9057	28936

Table 6.4. Distribution of number of hospitalization by asthma by age range in Catalonia at years 2004 and 2005

year	Age range		
	0-14	15-65	>65
2004	1113	2657	3046
2005	997	3942	4313

Table 6.5. Distribution of number of hospitalization by heart disease by age range in Catalonia at years 2004 and 2005

year	Age range		
	0-14	15-65	>65
2004	58	5203	10577
2005	10	5574	10911

The presented study was focused on children asthma exacerbation due to air pollution exposure. Relationship between hospital admissions in children population due asthma disease and air pollution exposure was studied by self-organizing maps feature extraction capabilities. Figures 6.3 and 6.4 present areal distribution of number of hospital admissions of children (0-14 years old) by asthma according municipal population (same

age range) at years 2004 and 2005 respectively. Visual inspection of Figures 6.3 and 6.4 and comparison with population distribution in Figure 6.2 reveals that high density areas like Barcelona metropolitan area (red color in Figure 6.2) do not exhibit high-frequency of hospital admissions by asthma in children population as it would be expected. It's important to mention that Barcelona is located in the Mediterranean coast and orography and climate conditions controls air pollution concentration distribution.

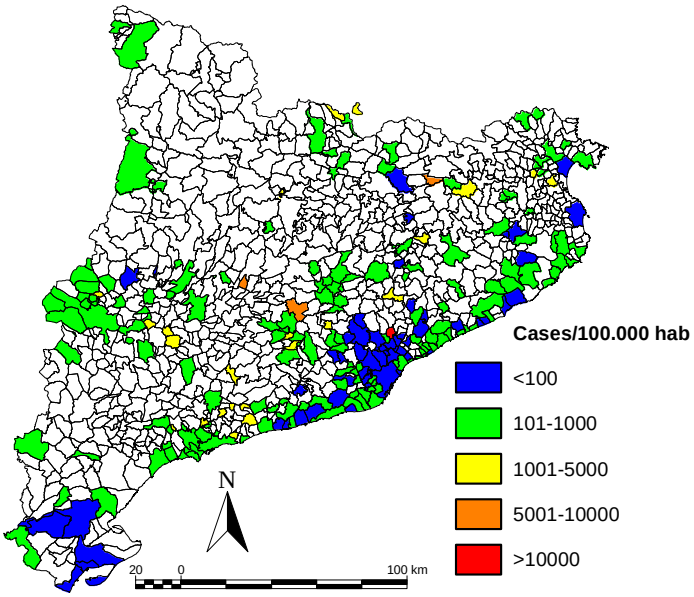


Figure 6.3. Distribution of number of hospital admissions by asthma in the age range 0-14 in Catalonia area for year 2004

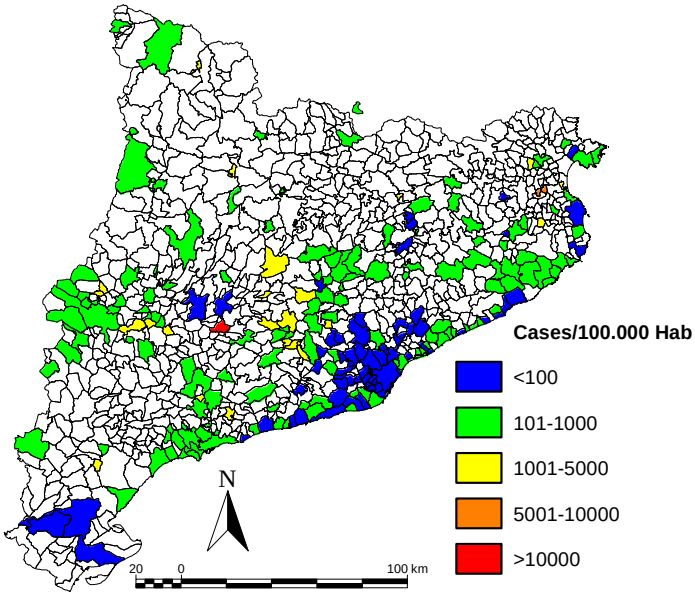


Figure 6.4. Distribution of number of hospital admissions by asthma in the age range 0-14 in Catalonia area for year 2005

Air concentration data for health-related pollutants like fine particulate matter (PM_{10}), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), benzene (C_6H_6) and oxide monoxide (CO) was obtained from Catalan environmental department for measurements stations presented in Figure 6.5. Tables 6.6 and 6.7 present the main statistics of pollutants mean annual concentrations in air for years 2004 and 2005. Nitrate and carbon dioxide exceeded the annual limit in several measurement stations. Table 6.8 presents maximum and minimum values of frequency of exceeding human health protective thresholds of health-related air pollutants (Table 6.1). Both particulate matter and ozone surpasses the regulatory values in all monitoring stations for the 24-h and 8-h measures respectively.

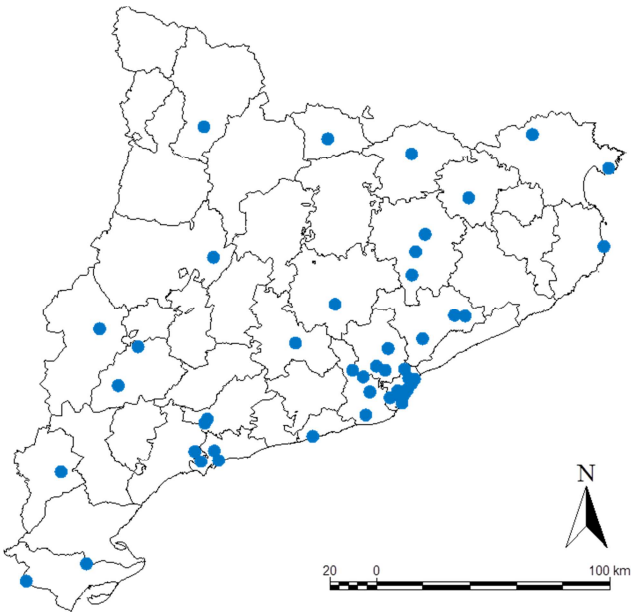


Figure 6.5. Air quality measurements stations considered in the HHRA for Catalonia area

Table 6.6. Main statistics of health–related air pollutants annual concentrations for year 2004

Pollutant	min ($\mu g/m^3$)	mean ($\mu g/m^3$)	max ($\mu g/m^3$)	std ($\mu g/m^3$)
PM_{10}	22.21	35.97	63.82	10.01
O_3	23.46	54.14	72.16	12.46
NO_2	15.74	23.13	49.22	7.33
SO_2	1.62	5.53	27.16	4.91

max: maximum value; mean: arithmetic mean; min: minimum value ; std: standard deviation

Table 6.7. Main statistics of health–related air pollutants annual concentrations for year 2005

Pollutant	min ($\mu g/m^3$)	mean ($\mu g/m^3$)	max ($\mu g/m^3$)	std ($\mu g/m^3$)
PM_{10}	18.63	31.32	62.35	8.01
O_3	26.47	58.77	78.22	12.14
NO_2	6.98	17.81	57.61	9.40
SO_2	1.43	4.40	16.66	2.57

max: maximum value; mean: arithmetic mean; min: minimum value ; std: standard deviation

Table 6.8. Minimum and maximum numbers of frequency of exceeding human health protective thresholds of health-related air pollutants at years 2004 and 2005

		2004		2005	
Pollutant	measure	min	max	min	max
PM_{10}	24-h	2	64	2	64
O_3	8-h	1	44	2	54
NO_2	1-h	0	6	0	11
SO_2	1-h	0	155	0	15
SO_2	24-h	0	10	0	1

max: maximum value; min: minimum value

6.3 Evaluation of SOM capabilities in HHRA

6.3.1 Concentrations maps using SOM

In previous chapters of this work, self-organizing maps (SOM) capabilities for areal interpolation had been explored and validated. In this section, SOMs were used to mimic cokriging techniques (Rallo, 2007) and produce smooth concentration maps for health-related air pollutants: PM_{10}, O_3, SO_2 and NO_2. C_6H_6 and CO were not considered in this section because of the lack of sufficient and reliable data in the study area.

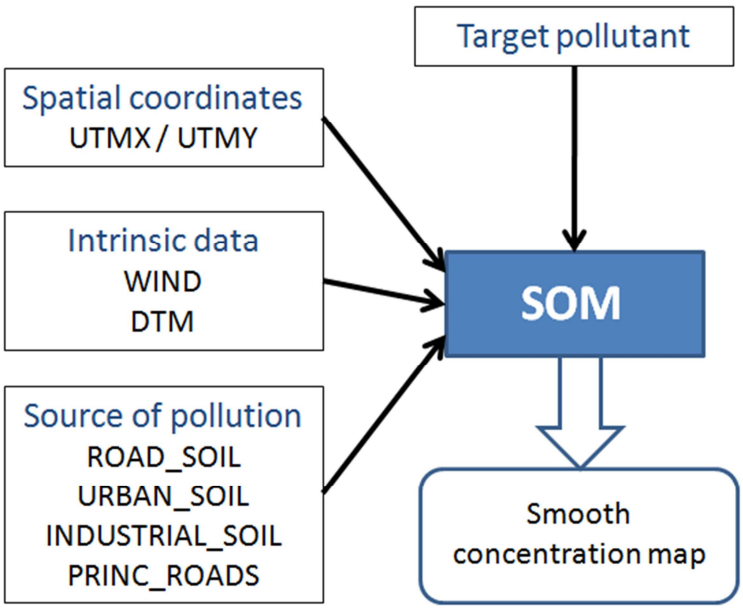


Figure 6.6. Training variables to produce pollutants concentrations maps using self-organizing maps

Several variables were identified and studied to generate spatial interpolated maps of the pollutants of interest. Input of SOM was composed of different variables aggregated in four categories presented in Figure 6.6: (i) spatial coordinates of the measurement station (UTMX and UTMY); (ii) media intrinsic data: wind direction (WIND) and topography from digital terrain model data (DTM); (iv) sources of pollution: percentage of road soil (ROAD_SOIL), percentage of urban soil (URBAN_SOIL), percentage of industrial soil (INDUSTRIAL_SOIL) and distance to principal roads (PRINC_ROADS); (iv) mean annual concentration of the target pollutant. Digital terrain model of Catalonia (Figure 3.2 of Chapter 3) was used to generate DTM variable. Wind variable was obtained from meteorological data from Catalan environmental agency. A buffer of 2-km diameter from measurements stations was applied to generate sources data features (percentages of urban, road and industrial soil) from land uses map of Catalonia (Figure 3.3 of Chapter 3). Distance to principal roads variable was calculated from principal roads map (Figure 4.2 of Chapter 4) by the MiraMon GIS software (Pons, 2006).

To illustrate the SOM cokriging interpolation, Figures 6.7 and 6.8 show ozone annual mean concentrations maps for years 2004 and 2005. C-planes of trained SOM are also presented and reveal the independency among training variables at both period of time.

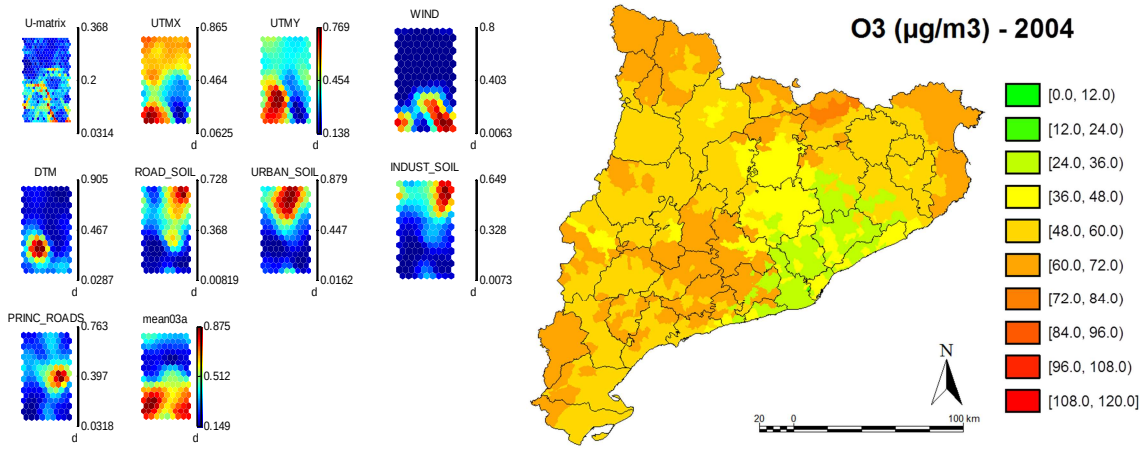


Figure 6.7. O_3 interpolation in Catalonia area for year 2004 by SOM (left) U-matrix and C-planes (right) SOM-O_3 concentrations map

Same procedure was applied to generate concentration maps for the other health-related pollutants at the two years in study. Analysis of PM_{10}, SO_2 and NO_2 concentrations maps generated by SOM interpolation capabilities also revealed independency of training variables and reliable areal distribution as in the case of ozone interpolated maps (Figures 6.7 and 6.8).

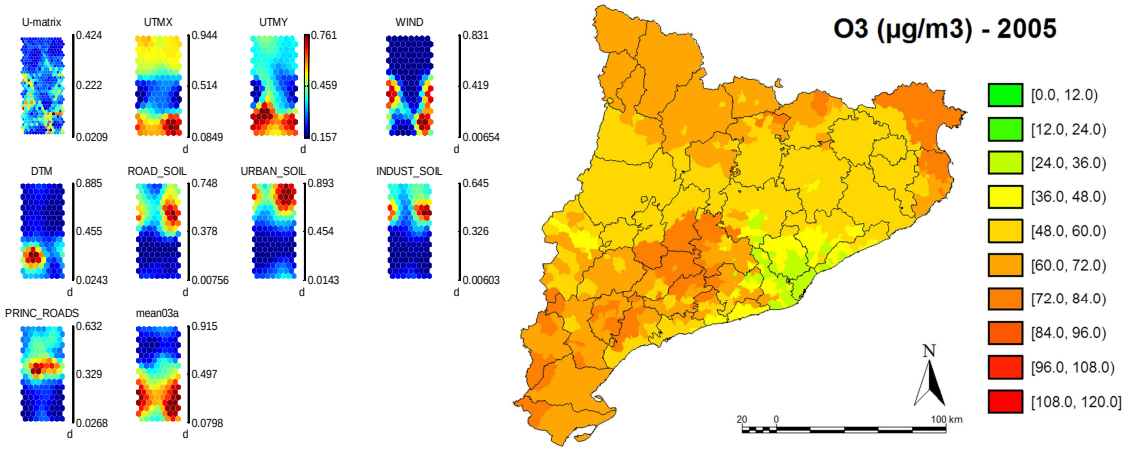


Figure 6.8. O_3 interpolation in Catalonia area for year 2005 by SOM (left) U-matrix and C-planes (right) SOM-O_3 concentrations map

6.3.2 Risk maps using SOM

In order to evaluate the capabilities of self-organizing maps in human health risk assessment, SOMs were used to generate risk maps related to asthma exacerbation in children (0-14 years old) and air pollution by self-organizing health-related pollutants data, as presented in Figure 6.9.

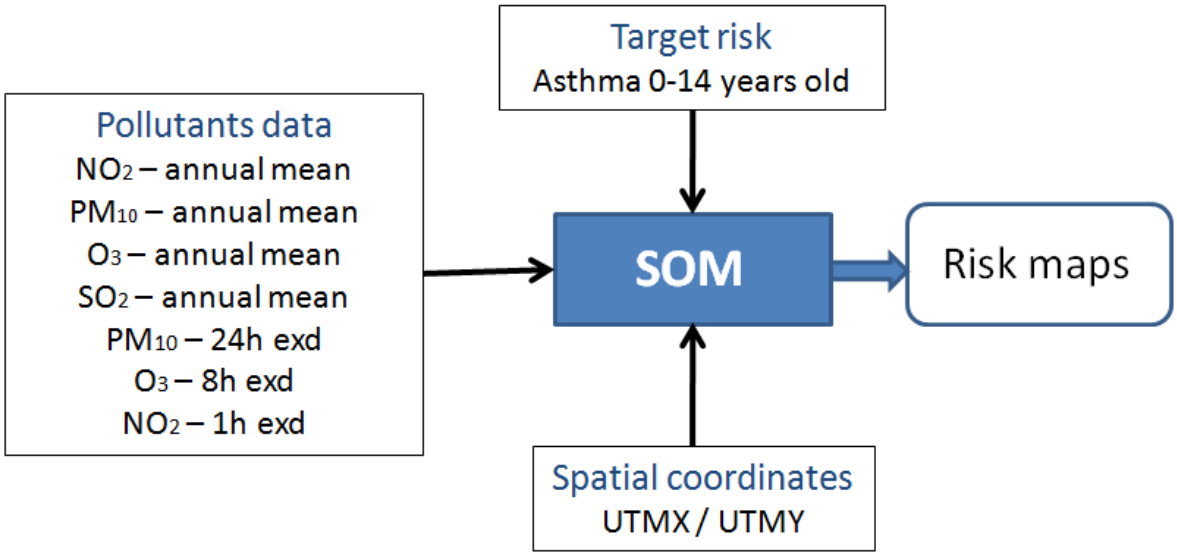


Figure 6.9. Scheme of SOM-based risk maps for asthma exacerbation due to air pollution in 0-14 years old population

Three major input variables were selected to perform SOM classification: (i) spatial coordinates (UTMX and UTMY); (ii) pollutants data (NO_2, PM_{10}, SO_2, O_3 annual mean; number of exceeding values of PM_{10} in 24-h, O_3 in 8h and NO_2 in 1h period); and (iii) target risk variable, in this case hospital admissions for asthma in children (0-14 years old).

Toroidal SOMs were selected and different SOMs sizes were evaluated using Matlab´s SOM Toolbox. As described in previous chapters, a *mask* for grouped input variables in the training process was used in order to give same weights to each of the three categories

Risk maps of asthma in children for Catalonia were obtained using SOM and presented in Figures 6.10 and 6.11 for years 2004 and 2005 respectively. The risk maps labels were selected using the Euclidean norm of centroids of SOM's clusters and sorting for the final classification. This approach might not be appropriate for all target variables and should be studied and reformulated to generate more reliable risk labels.

Visual inspection of risk maps of Figures 6.10 and 6.11 reveals differences in the risk estimations (labeling) from year 2004 to 2005. This effect it's a consequence of the time variability in the pollutants concentrations maps used to generate risk maps.

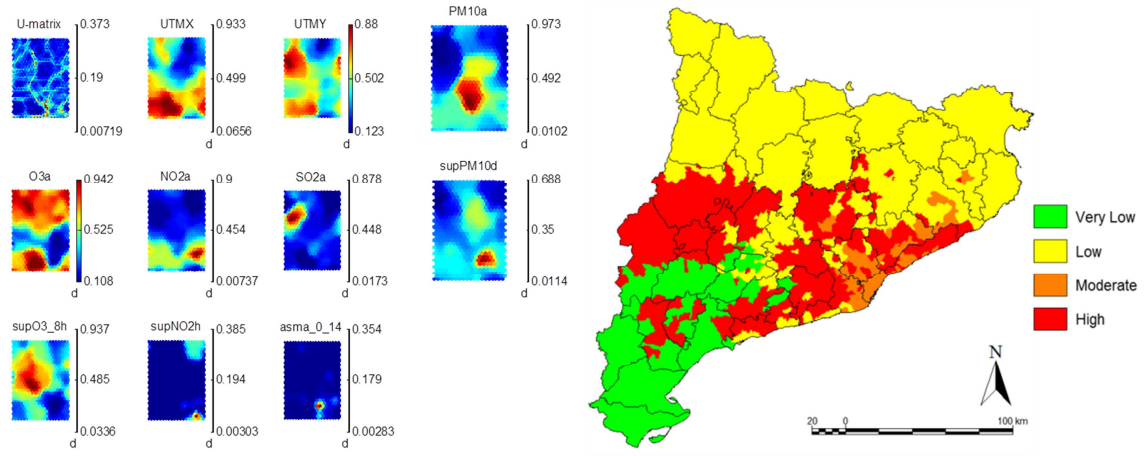


Figure 6.10. Children asthma attacks hospital admissions and air pollution relationship in Catalonia area for year 2004 (left) U-matrix and C-planes (right) SOM risk map model

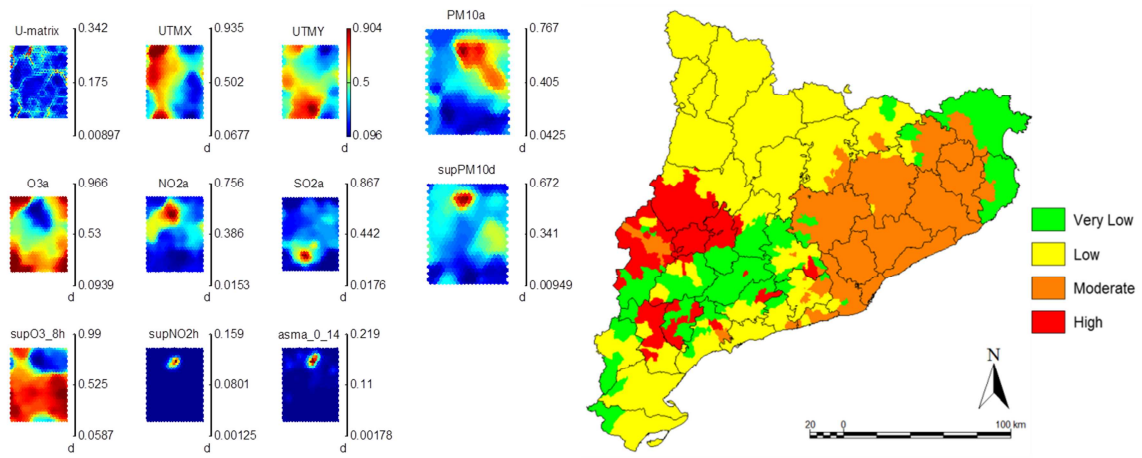


Figure 6.11. Children asthma attacks hospital admissions and air pollution relationship in Catalonia area for year 2005 (left) U-matrix and C-planes (right) SOM risk map model

6.4 Conclusions

The use of self-organizing maps for clustering variables have been demonstrated elsewhere and validated in the present work in the area of human health risk assessment. The produced SOM-based risk maps are not intended to be used as reference of an epidemiological study because they only reveal some cause-effect relationships between selected air pollutants and asthma in children based in hospital admissions at residence of patients during a two-year period. Application of SOM clustering methodology in an extended period of time is necessary to obtain more reliable results.

Definition of a SOM-based risk index should be evaluated in order to improve the labeling of risk classes and formulation of the index as presented for groundwater vulnerability in this work. Risk target variable could be used to evaluate the resulting risk index.

Self-organizing maps demonstrated their capabilities for extracting cause-effect relationships in human health effects due the cumulative exposure or different air pollutants. A case study of hospital admissions for asthma in children (up to 14 years old) in Catalonia area during years 2004-2005 revealed important relationships with air pollutants using SOM as non-linear classifier.

6.5 References

BELL M.L., K. EBISU, R.D. PENG, J.M. SAMET and F. DOMINICI (2009) "Hospital admissions and chemical composition of fine particle air pollution". American Journal of Respiratory and Critical Care Medicine 179: 1115-1120.

CHANG C., S. TSAI and C. YANG (2007) "Air pollution and hospital admissions for cardiovascular disease in Tapei, Taiwan". Epidemiology 18(5): S8-S9.

CHRISTAKOS, G., R. A. OLEA and H-L.YU (2007) "Recent results on the spatiotemporal modelling and comparative analysis of Black Death and bubonic plague epidemics". Public Health 121: 700-720.

CLARK N.A., P.A. DEMERS, C.J. KARR, M. KOEHOORN, C. LENCAR, L. TAMBURIC and M. BRAUER (2010) "Effect of early life exposure to air pollution on development of childhood asthma". Environmental Health Perspectives 118(2): 284-290.

DELFINO R.J., N. STAIMER, T. TJOA, D. GILLEN, M.T. KLEINMAN and S. CONSTANTINOS (2008) "Personal and ambient air pollution exposures and lung function decrements in children with asthma". Environmental Health Perspectives 116(4): 550-558.

DELFINO R.J., J. CHANG, J. WU, C. REN, T. TJOA, B. NICKERSON, D. COOPER and D.L. GILLEN (2009) "Repeated hospital encounters for asthma in children and exposure to traffic-related air pollution near the home". Annals of Allergy, Asthma & Immunology 102(2): 138-144.

DOCKERY, D., and A. POPE (1996) "Epidemiology of acute health effects: Summary of time-series studies". In: Particles in our air: concentration and health effects. Wilson R, Spengler JD (eds), Cambridge, MA. USA, Harvard University Press, pp. 123-147.

DOMINICI F., R.D. PENG, M.L. BELL, L. PHAM, A. MCDERMOTT, S.L. ZEGER and J.M. SAMET (2006) "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases". Journal of the American Medical Association 295(10): 1127-1134.

DOMINICI, F., R.D. PENG, C.D. BARR and M.L. BELL (2010) "Protecting human health from air pollution: Shifting from a single-pollutant to a multipollutant approach". Epidemiology 21(2): 187-194.

E.E.A. (1998) "Environmental Risk Assessment – Approaches, Experiences and Information Sources". European Environmental Agency Publication.

GEHRING U., A.H. WIJGA, M. BRAUER, P. FISCHER, JC. DE JONGSTE, M. KERKHOF, M. OLDENWENING, H.A. SMIT and B. BRUNEKREEF (2010) "Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life". American Journal of Respiratory and Critical Care Medicine 181: 596-603.

GÖTSCHI, T., J. HEINRICH, J. SUNYER and N. KÜNZLI (2008) "Long-term effects of ambient air pollution on lung function. A review". Epidemiology 19(5): 690-701.

HALONEN J.I., T. LANKI, T. YLI-TUOMI, M. KULMALA, P. TIITTANEN and J. PEKKANEN (2008) "Urban air pollution, and asthma and COPD hospital emergency room visits". Thorax 63: 635-641.

JERRETT M., K. SHANKARDASS, K. BERHANE, W.J. GAUDERMAN, N. KÜNZLI, E. AVOL, F. GILLILAND, F. LURMAN, J.N. MOLITOR, J.T. MOLITOR, D.C. THOMAS, J. PETERS and R; MCCONNELL (2008). "Traffic related air pollution and asthma onset in children: a prospective cohort study with individual exposure measurement". Environmental Health Perspectives 116(10): 1433-1438.

KAMPA, M. and E. CASTANAS (2008) "Human health effects of air pollution". Environmental Pollution 151(2):362-367.

LINPING C., K.L. MENGERSEN and S. TONG (2007) "Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia". Science of the Total Environment 373(1): 57-67.

O'CONNORS G.T., L. NEAS, B. VAUGHN, M. KATTAN, H. MITCHELL, E.F. CRAIN, R. EVANS, R. GRUCHALLA, W. MORGAN, J. STOUT, G.K. ADAMS and M. LIPPMANN (2008) "Acute respiratory health effects of air pollution in children with asthma in US inner cities". Journal of Allergy and Clinical Immunology 121(5): 1133-1139.e1.

PONS, X. (2006) "MiraMon Geographic Information System and Remote sensing software". http://www.creaf.uab.es/miramon/

RALLO, R. (2007) "Multi-tier framework for the inferential measurement and data-driven modeling". PhD dissertation. Universitat Rovira i Virgili. Spain.

SAMOLI E., P.T. NASTOS, A.G. PALIATSOS, K. KATSOUYANNI and K.N. PRIFTIS (2011) "Acute effects of air pollution on pediatric asthma exacerbation: Evidence of association and effect modification" Environmental Research 111: 418-424.

SCHWARTZ J. and R. MORRIS (1995) "Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan". American Journal of Epidemiology 142(1): 23-35.

VARDOULAKIS, S. and P. KASSOMENOS (2008) “Sources and factors affecting PM_{10} levels in two European cities: Implications for local air quality management”. Atmospheric Environment 42(17): 3949-3963.

W.H.O. (2003) “Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide”. World Health Organization, Regional Office for Europe, pp. 98.

Chapter 7

Conclusions

7.1 Main conclusions

The main conclusion of this thesis work is that artificial neural networks (ANN) can address environmental risk assessment (ERA) in different ecological receptors such as groundwater, air or human health over the study area of Catalonia Mediterranean region. Four scenarios were studied to evaluate the applicability of data-driven ERA-ANN based approach.

The Self-Organizing Map (SOM) algorithm provides a consistent framework that can be successfully applied at different levels of the ERA framework and at different spatial resolutions. The proposed methodology for groundwater vulnerability assessment has been tested at two different scales. At the local level in the Camp de Tarragona hydrogeological unit, where the concentration distribution functions of diverse stressors have been determined from available data, the clustering capabilities of the SOM provide vulnerability estimates without requiring previous expert's knowledge ratings of numerical variables.

At wider scales such as the regional level of Catalonia, the SOM has been successfully used to deal with missing data, spatial interpolation, probabilistic risk analysis, and intrinsic and specific vulnerability assessment. This latter aspect is addressed by the development of a new vulnerability index which is independent of expert's criteria and is able to integrate several sources of diverse hydrogeological and climatic information.

The new vulnerability index (vIndex) is standardized and discretized in such a way that it is independent of the scale of the geographical region considered. The values obtained from this new index can thus be easily applied to estimate groundwater vulnerability at different scales within Europe. The new methodology provides more detailed and geographically consistent vulnerability maps than the well-established DRASTIC methodology in all its variations proposed previously in the literature.

The SOM-based groundwater vulnerability approach can help regulators and policy makers to understand the relationships between the potential stressors of concern in an environmental risk scenario such as the pollution of groundwater and the vulnerability of

drinking water sources. As the proper management of water resources is becoming a major concern in Europe and in the rest of the world, the proposed SOM based approach for groundwater intrinsic and specific vulnerability assessment provides a reliable and adaptable tool for resource planning and decision making.

Integrated assessment of anthropogenic sources and ecological risks of lead pollution was achieved by gathering georeferenced information and pollution data and applying this novel methodology using a Fuzzy ARTMAP (FAM) neural classifier. This part of the study demonstrated that Fuzzy ARTMAP neural classifier is able to establish multi-variable cluster analysis with cause-effects relationships relevant for exposure and human health such as lead exposure assessment in groundwater. This methodology can be applied to different pollutants in groundwater using proper identification of anthropogenic sources and hydrogeological behavior of the contaminant. Also, cumulative assessment can be accomplished by studying the cause-effect relationships of many pollutants in a specific area or ecological receptor.

Spatio-temporal air quality assessment was achieved by the generation of smooth maps of PM_{10} annual concentration using two different techniques: Bayesian Maximum Entropy (BME) and Temporal Self-organizing maps (T-SOM). Both BME (well-known technique) and T-SOM approaches (novel technique) are capable of identifying hotspot areas in annual PM_{10} concentration maps by interpolating spatio-temporal PM10 daily data at monitoring stations over Catalonia.

In the BME approach the use of “soft” data, such as the daily PM_{10} concentration distribution, improves the resulting estimations since it includes temporal data dispersion in the interpolation process (specific knowledge). The need of a general knowledge component, which requires the fitting of a model for the experimental covariance characterization, is a cost-consuming task. T-SOM has demonstrated to be a very powerful tool to generate annual spatio-temporal maps by extracting relevant information of spatio-temporal daily data (only specific knowledge), without the need of covariance generation.

Finally, self-organizing maps demonstrated their capabilities for extracting cause-effect relationships in human health effects due the cumulative exposure or different air pollutants. A case study of asthma emergency consultations in children (1 to 14 years old) in Catalonia area during years 2004-2005 revealed important relationships with air pollutants using SOM as non-linear classifier.

7.2 Summary of data-driven ANNmethodology in ERA

Artificial neural networks have been demonstrated their capabilities in data-driven modeling of diverse complex engineering problems. This thesis contributed to enhance ANN applicability to environmental risk assessment in different media, spatial and temporal scenarios.

As the final contribution of this research work in the ERA framework is summarized in a general methodology to properly apply ANN in data-driven modeling and cause-effect relationships. Figure 7.1 presents the general schematic methodology of data-driven ANN in ERA.

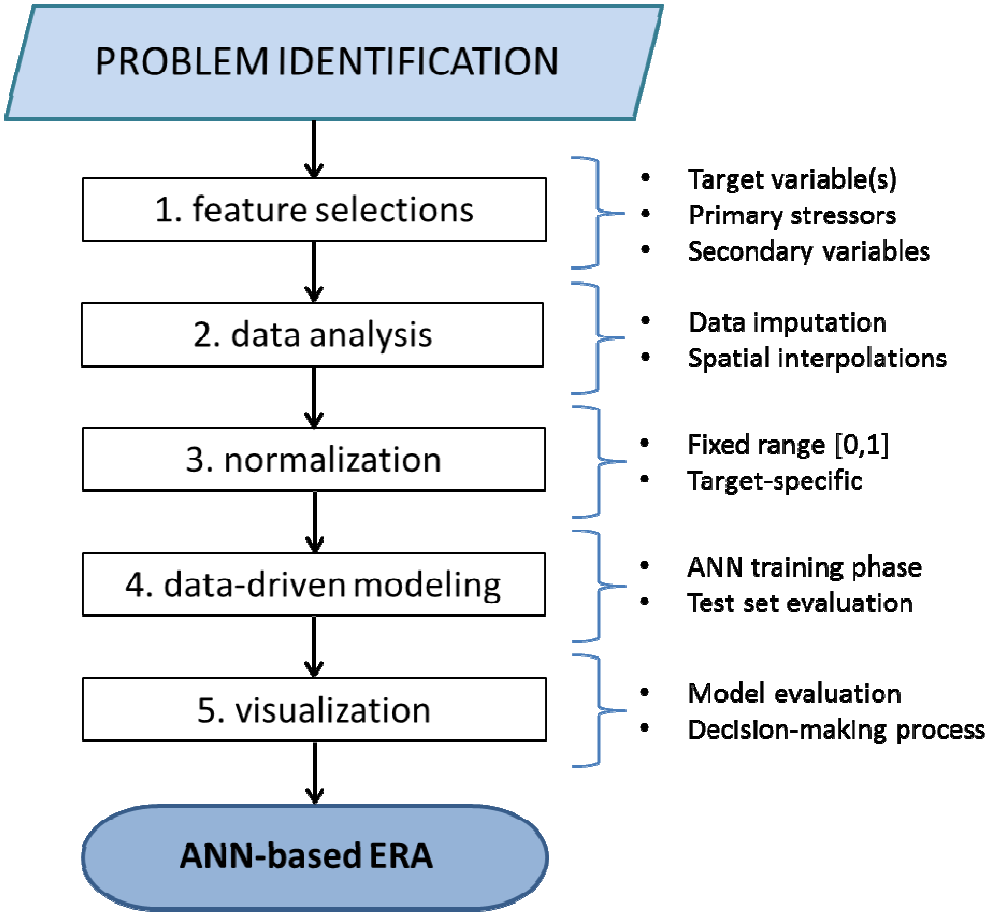


Figure 7.1. General schematic methodology of data-driven ANN in ERA

7.3 Research opportunities

Future extensions of the use of self-organizing maps for intrinsic groundwater vulnerability assessment should provide mechanisms for automatic tuning the optimal neighborhood used for exposure modeling. Also, the categorization of the vulnerability classes could be exploited to produce vulnerability-driven risk assessment models in which accurate probabilistic risk models could be adjusted to each vulnerability category. Additionally, the use of hierarchical ensembles of SOMs could provide an integrated view of vulnerability at different spatial scales and facilitate the inference of relationships between vulnerability estimates at each scale.

The performance of Fuzzy ARTMAP neural network for lead exposure in Catalonia area could be improved by incorporating new variables in the analysis, to account, for example, soil's degradation capacity or climatologic variations through the area of study. Spatio-temporal analysis should also be done, by gathering data in an expanded time period. Also, and use of SOM as a prediction tool could be evaluated for lead assessment in groundwater.

Future work in Temporal-SOM approachfor air quality assessment would be the inclusion co-variables in the T-SOM training process to improve the accuracy of the spatio-temporal estimations, mainly in unmonitored areas. Some variables to consider might be wind direction, average daily traffic, industrial pollutant sources, and cultivated lands, among others. The final step will be characterization of particulate matter sources using co-variables in T-SOM in order to find non-linear relationships between PM_{10} measurements and different pollution sources. Exploratory analysis of PM_{10} data in Catalonia evidences important polluted areas around the cities of Barcelona and Tarragona. Also the PM_{10} monitoring network has to be evaluated and redesigned in order to better represent the atmospheric quality status over the whole area. Because of the clustering capabilities of the SOM, it could be used to design a new and optimized monitoring network that take into account the climatologic variability and stressors sources through Catalonia area.

In the application of SOM in human health risk assessment, definition of a SOM-based risk index should be evaluated in order to improve the labeling of risk classes and formulation of the index as presented for groundwater vulnerability in this work. A risk target variable could be used to evaluate the resulting risk index. Also, the application of SOM clustering methodology in an extended period of time is necessary to obtain more reliable results.

Annex A

Research Contributions

A.1 Paper on DRASTIC-SOM for Camp de Tarragona published in Environmental Modelling & Assessment

Environ Model Assess (2011) 16:119–133
DOI 10.1007/s10666-010-9230-6

Mapping Cumulative Environmental Risks: Examples from the EU NoMiracle Project

Alberto Pistocchi · Jan Groenwold · Joost Lahr · Mark Loos · Marelys Mujica · Ad M. J. Ragas · Robert Rallo · Serenella Sala · Uwe Schlink · Kathrin Strebel · Marco Vighi · Pilar Vizcaino

Received: 30 October 2009 / Accepted: 27 May 2010 / Published online: 9 July 2010

Abstract We present examples of cumulative chemical risk mapping methods developed within the NoMiracle project. The different examples illustrate the application of the concentration addition (CA) approach to pesticides at different scale, the integration in space of cumulative risks to individual organisms under the CA assumption, and two techniques to (1) integrate risks using data-driven, parametric statistical methods, and (2) cluster together areas with similar occurrence of different risk factors, respectively. The examples are used to discuss some general issues, particularly on the conventional nature of cumulative risk maps, and may provide some suggestions for the practice of cumulative risk mapping.

Keywords Cumulative environmental risk · GIS mapping · Mixtures · Multiple stressors · Pesticides · Metals · Spatial distribution

1 Introduction

Over the last few years, there has been an increasing interest in the representation of chemical risks for purposes such as decision support, risk communication, and scientific research. As mapping technologies become more and more routinely used, attention for the underlying methods of risk mapping and visualization is growing (e.g., [2, 13,

Electronic supplementary material The online version of this article (doi:10.1007/s10666-010-9230-6) contains supplementary material, which is available to authorized users.

A. Pistocchi · P. Vizcaino
European Commission Joint Research Centre,
Via E.Fermi,
1-21027 Ispra, VA, Italy

J. Groenwold · J. Lahr
Alterra, Wageningen UR,
P. O. Box 46, 6700 AA Wageningen, The Netherlands

M. Loos · A. M. J. Ragas
Department of Environmental Science,
Institute for Water and Wetland Research,
Radboud University Nijmegen,
Heyendaalseweg 135,
6525 AJ Nijmegen, The Netherlands

M. Mujica · R. Rallo
Departament d'Enginyeria Química i Departament d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili,
Av. Paisos Catalans, 26,
43007 Tarragona, Catalonia, Spain

S. Sala · M. Vighi
Department of Environmental Sciences,
Universita' di Milano Bicocca,
Piazza della Scienza 1,
20126 Milan, Italy

U. Schlink · K. Strebel
Helmholz Centre for Environmental Research—UFZ,
Permoserstrasse 15,
04318 Leipzig, Germany

Present Address:
A. Pistocchi (✉)
European Academy,
Viale Druso,
1-39100 Bolzano, Italy
e-mail: alberto.pistocchi@eurac.edu

20, 31, 60, 61]). Lahr and Kooistra [28, 29] recently categorized the different types of risk maps that exist and reviewed the methods to make them. They distinguish, among others, between maps of contamination, (potential) exposure, vulnerability, and "true risk" for single or multiple stressors.

In general, risk is determined by the concurrence of chemical pollution and vulnerable receptors (e.g., organisms, populations, communities, ecosystems, ecosystem resources, and services). Both can be mapped. In the past, pollutant concentrations were often used as a proxy for risk on maps. It was implicitly assumed that vulnerable receptors were homogeneously distributed over the analysis area, generally owing to the lack of information on the spatial distribution of these receptors. However, receptor vulnerability is increasingly being included in spatial analyses of risk (e.g., [10, 21, 32, 49]). Maps that show the spatial distribution of vulnerable receptors are called vulnerability maps, e.g., ground water vulnerability maps (e.g., [47, 62]).

True risk can be defined as "the probability of an adverse effect on man or the environment resulting from a given exposure to a chemical or mixture" [53]. True risk can be calculated from the combination of exposure and receptor vulnerability. The mapping of true risk caused by a single pollutant is conceptually straightforward, although practical problems are posed by lacking or incomplete data, the consideration of multiple exposure pathways, and the mobility of receptors. Additional problems may arise when considering cumulative risks, i.e., those arising from multiple chemicals acting as a mixture or chemicals together with other (non-chemical) stressors such as aridity, climate, or land use change. Risks may be posed to single as well as multiple receptors, and exposure of (mobile) receptors may vary between different micro-environments.

Lahr and Kooistra [28, 29] discuss the most important issues in risk mapping and provide some general rules of thumb for making environmental risk maps for communication purposes. One of the limitations they identify is that only one or very few parameters can be represented on a single map. This particularly poses problems for cumulative risk maps which by definition have to deal with multiple parameters, and are increasingly a subject of interest in chemical pollution management (e.g., [22, 54]). The most applied solution to this problem is to express the overall risk in terms of a single indicator and to map the outcome, although examples exist of maps that visualize multiple parameters simultaneously [28, 29]. Despite the interest in cumulative risk mapping, little guidance exists about which methods to adopt in different circumstances. Spatially distributed chemical risk assessment remains a conceptually complex procedure, although tools for spatially explicit modeling are increasingly available and attractive (see, e.g., the discussion in [38]). As a complement to Lahr and Kooistra's critical review, we present here a range of methods for the analysis and presentation of cumulative risks which were recently developed and applied within the European NoMiracle project (http://nomiracle.jrc.ec.europa.eu). By comparing the different methods, we aim at providing the reader with some general insights and guidelines for analyzing and mapping cumulative risks.

We first identify a range of appropriate methods for cumulative risk assessment. These include models of mixture toxicity, models of variable exposure, data-driven risk mapping, and classification (or clustering) techniques based on known risk factors. Then, the different methods are presented one by one through examples. We finally propose summary considerations, which may help practitioners in need of mapping cumulative chemical risks. The methods we deal with, and the examples we use, are summarized in Table 1.

2 Materials and Methods

A widely used scheme to characterize the combined action of multiple chemical substances is that of Bliss [5], further developed by Plackett and Hewlett [39]. Stemming from that scheme, two different approaches to modeling mixture toxicity are typically used, the concentration addition (CA) and the independent action (IA) models [17]. The CA approach assumes that different chemicals act together as their respective sum. Concentrations should be added up just after appropriate normalization. One way to do so is to divide them by a comparable threshold concentration, such as the widely used 50% effect concentration (EC_{50}). The IA approach assumes that the overall response of an ecosystem to a mixture of chemicals is the sum of responses to individual chemicals. The two models are applicable to chemicals with the same mode of action or to chemicals with different modes of action respectively.

Assuming that environmentally relevant mixtures have heterogeneous mechanisms of action, a two-stage prediction approach (TSP) was developed [24] by combining the CA and IA models. Conceptually, the TSP approach is the best to assess pesticide mixtures that can be expected to be neither strictly similarly nor strictly dissimilarly acting. Taking into account that the mixture responses calculated using the CA model are usually higher than those calculated with the IA model, CA can be assumed as a conservative but "reasonable" worst case [6, 12, 14, 16, 24].

The metric of potential ecotoxic risk for chemical mixtures under the CA assumption are the toxic units (TU) of the mixture:

$$\mathrm{TU}_m = \sum_{i=1}^{n} \mathrm{TU}_i = \sum_{i=1}^{n} \frac{C_i}{\mathrm{EC}_{x,i}} \quad (1)$$

Table 1 Summary of the features of the methods presented in the paper

Method	Cumulative aspect	Example presented	Underlying approach	Target of risk	Measure of risk presented in the map
Method 1: models of mixture toxicity	Mixture of chemicals acting together	Mixtures of pesticides at European, national, and regional scale	Deterministic calculation of PECs with GIS; concentration addition model	Ecosystems (generic)	Toxic units. Potential exposure
Method 2: variable exposure modeling	Exposure of individuals to different chemicals unevenly distributed in space	Exposure to heavy metals related to foraging behavior in a Dutch floodplain	Receptor-oriented agent-based modeling	Ecosystems (specific organisms)	Hazard quotient for individual organisms. "True risk"
Method 3: data-driven risk mapping	Pollution from a mixture of sources	Ambient air-borne benzene in Leipzig	Regression involving risk factors and spatial auto-correlation	Human health	Chemical concentration. Contamination
Method 4: classification based on known (a priori) risk factors	Different factors concurring to natural resource vulnerability	Aquifer vulnerability in Catalonia	SOM classification	Aquifer	Vulnerability

where C_i is the actual concentration of the individual chemical "*i*" in the mixture, $EC_{x,i}$ is the ecotoxicological endpoint (e.g., EC_{50}) of the individual chemical *i*, and TU_i are the toxic units of the individual chemical *i*, i.e., the fraction of the ecotoxicological endpoint produced by the individual chemical *i* ($TU_i = C_i/EC_{x,i}$). Toxic units are also called, equivalently, exposure–toxicity ratio (ETR) or hazard quotient (HQ).

The CA approach is widely used in the assessment of pesticide risk in Europe. The HAIR project (http://www.rivm.nl/rvs/risbeoor/Modellen/HAIR.jsp), among others, has proposed risk indicators related to terrestrial and aquatic organisms, which are summarized in the supporting information (SI, Table S1). These indicators involve mathematical models which may be relatively complex and require a number of input parameters. However, it can be easily shown that all indicators ground on the CA assumption, and therefore, they are linearly related to a limited number of parameters representing pesticide pressure. In some cases, concentrations are not needed explicitly in order to predict impacts on a specific endpoint, as they are linearly related to other variables more easily predicted, such as emissions and wind drift or loads to aquatic ecosystems.

We present three applications of the CA model to pesticides at different spatial scales: a river catchment (within the region of Lombardia, Italy), a country (The Netherlands), and continental Europe.

In such examples, as very often in spatial risk assessment practice, receptor spatial distribution and vulnerability is assumed to be uniform; therefore, risk is directly represented by a function of concentrations, given by the mixture toxicity model adopted: These examples may be considered in between "potential exposure" and true risk following [28, 29].

Sometimes, inhomogeneous risk from chemicals arises from different characteristics in space of individuals or cohorts exposed, which requires modeling explicitly the behavior of receptors. A typical example in this sense is the consideration of a different intake by adults and children in protocols of human risk assessment. Particularly, when exposed to multiple stressors, the activity pattern of the receptor plays an important role: At different locations and moments in time, receptors are exposed to varying combinations and concentrations of multiple stressors [40]. Therefore, exposure models for multiple stressors should primarily focus on the receptor, and not on the stressor(s), as in the case of wildlife exposure to metals according to the respective foraging habits, that we present as an example. This case can be regarded as in between an exposure mapping and a true risk mapping according to [28, 29].

When heterogeneous factors concur to produce risk, and an explicit model of their interaction cannot be defined, a

"data driven" approach should be adopted. One example is provided about mapping risk due to benzene in air in Leipzig, Germany. In that case, a risk indicator is defined (benzene concentration), and a statistical model involving the most relevant explanatory factors is applied to interpolate point measurements to a continuous representation of the risk indicator. For the specific example, we assume concentration as a proxy of risk, therefore neglecting the variability of receptor conditions. According to Lahr and Kooistra [28, 29], then this is an example of a contamination map.

When no information on actual impacts is available, one may still combine different risk factors based on prior knowledge. Combinations may be rule-based classifications, or alternatively ground on formal clustering techniques. Among the latter, the Self Organizing Map (SOM) technique [27] is an unsupervised neural network algorithm that projects (classifies) high-dimensional data into a two or three-dimensional grid of units (clusters) while preserving the original topology of the input space and facilitating the visualization of hidden patterns present in the data [27]. SOM units are organized on a regular hexagonal lattice that defines the neighboring structure of the map units. The algorithm is based on competitive learning [25, 26, 58] where units gradually become sensitive to different input categories of the input space.

The SOM is a powerful clustering tool that has demonstrated to be appropriate for the classification and visualization of complex datasets including highly nonlinear relationships. The component planes (C-planes) are the most important analysis and visualization tools since they provide the distribution over the map of the values corresponding to each component of the input data vectors. Straightforward correlations and relationships in the input dataset can be found by simultaneously comparing several C-planes [57]. The clustering structure of the input space is visualized using the unified distance matrix (U-matrix), which is constructed by measuring the distances between all units in the map. The U-matrix is usually post-processed by clustering its components to produce coarser data partitions. We present an application of this technique through the example of aquifer vulnerability to pollution in Catalonia, Spain.

3 Examples of Cumulative Environmental Risk Mapping

3.1 Concentration Addition

As a *first example*, we refer to the distribution of agricultural pesticides in Europe. Screening level maps of pesticide mass in soil and load to streams in Europe are available [35, 37]. These maps are linearly related to predicted environmental concentrations as used in the HAIR indicators and therefore can be directly expressed using the CA concept. The predicted mass in soil and load to streams for each substance class, represented each time by its "most dangerous chemical," has been used in a weighted summation, so to express, in terms of toxic equivalents to one substance, total mass and load as a cumulative risk indicator for terrestrial and aquatic ecosystems.

Unfortunately, currently available data on pesticides for Europe are limited to chemical substance classes and not to individual pesticides within each class. Therefore, for the sake of illustration, we assumed that each chemical substance class is composed of the most dangerous chemical of the class, selected as the one having the lowest toxicity threshold within its class. We retrieved physico-chemical and toxicological properties for the active substances within each class from the FOOTPRINT online Pesticide Properties Data Base (PPDB) (www.eu-footprint.eu). The weights of the generic j-th chemical, v_j and w_j, used to sum together mass in soil and load to streams for different chemicals can be estimated as:

$$v_j = \frac{T_{t,j}}{\max_{j\in\{1:nc\}}(T_{t,j})} \qquad w_j = \frac{T_{a,j}}{\max_{j\in\{1:nc\}}(T_{a,j})} \tag{2}$$

where $T_{x,j}$ (x=t, a) is the toxicity threshold (no observed effect concentration (NOEC), 50% lethal concentration (LC_{50}), or similar metrics) of chemical j, for terrestrial and aquatic ecosystems, respectively, while nc is the number of chemicals considered.

Unfortunately, not for all active substances toxicity data are provided in the database. Terrestrial endpoints are better covered than aquatic ones. Bees represent an endpoint for spray drift only, as they are impacted only by pesticide reaching non-target vegetation and crops. For soil ecology, earthworms are a more representative endpoint. However, chronic toxicity data for earthworms are far less abundant than acute LC_{50}. Table S2 in the SI shows the percentage of pesticides with toxicity data available, according to the five most common endpoints tested. Although absolute toxicities vary depending on the receptor under consideration, and the temporal span of exposure (acute, chronic), in the absence of more detailed information sometimes, it is assumed that the relative chronic toxicity of substances is reflected by the relative acute toxicity (e.g., [11]).

In the present application, we consider acute toxicity to earthworms for terrestrial organisms, and NOEC at 21 days for aquatic organisms. Table S3 in the SI indicates the most dangerous chemical selected for each chemical class, for terrestrial and aquatic ecosystems, respectively. The same toxicity data are used to derive weights with Eq. 2, also provided in the same table.

The proposed assessment prospects an indicator of pesticide risk in Europe for aquatic and terrestrial ecosystems for different years, allowing an estimate of trends in the overall risk related to pesticides. The results of these calculations for the year 2003 are shown in Figs. 1 and S1 of the SI, respectively. The two maps highlight the cumulative spatial distribution of pesticides, with reference to two different endpoints (terrestrial and aquatic), taking as reference substances the most toxic ones in the two cases (picoxystrobine, a strobilurine fungicide, and omethoate, an organophosphorus insecticide, respectively). The spatial distribution may be grossly similar, but significant differences arise due to variations in physico-chemical properties of the substances, hence the different weighting for the two endpoints.

Figures 1 and S1 of the SI highlight potential problem areas or "hot spots" in a specific year (2003). In general, the use of pesticides in Europe is rather widespread; hot spots are predicted in Spain, Italy, France, and The Netherlands. Some countries (like Poland) show extensive presence of medium–high levels. Often hot spots are associated with vineyards, generally bearing the highest pesticide application from the 20 classes considered in this study.

A comparison of maps as shown in Figure S1 of the SI, and Fig. 1, with the corresponding ones for the year 1992 (not shown here for simplicity), yields a picture of the variation in overall risk due to pesticides, with reference to a specific receptor (Fig. 2).

As it appears, the situation in Europe resulting from the example calculation is rather variable in space, showing areas of increase and decrease of overall pressure (load, mass).

In the case of mass in soils, representing pressure on terrestrial ecosystems, there is generally a persistence or increase between 1992 and 2003, while on aquatic ecosystems, there is a general decreasing trend. However, in both cases, important differences arise across regions. The different behavior is linked to the differences in trends in the use of the different substances contributing to the overall toxicity to terrestrial and aquatic endpoints.

It is important to stress that the maps provide the maximum toxic equivalent for a given combination of emissions from the different chemical substance classes. Therefore, the result is in general an upper limit of the cumulative toxicity. The only way to obtain an estimate of the actual cumulative toxic equivalent would be through referring emissions to individual chemicals, of which on the other hand use data are not yet available for Europe. Also, the spatial distribution must be regarded as a realization of the random field of pesticide use, as there is not sufficient information to allocate to each country and crop within a country the corresponding pesticide use. The method just outlined is applied to map the overall impact of pesticides on terrestrial and aquatic organisms in Europe. A similar approach can be extended to other risk indicators, and particularly to human health risk; this is anyway beyond the scope of the example presented here.

When the assessment does not concern general trends only, but requires higher realism, the temporal as well as spatial distribution of pesticide emissions needs to be taken into account to reflect realistic field conditions.

A *second example* deals with region-specific risks of pesticide mixtures [41]. It relates to water bodies in a pilot

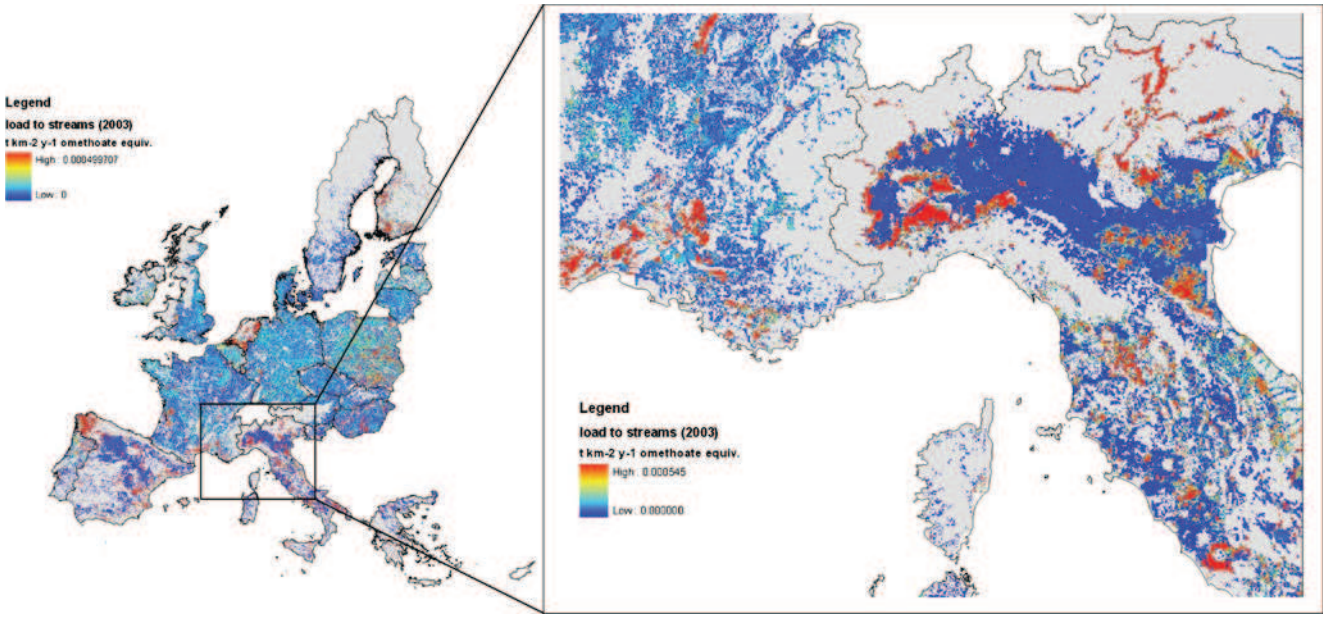


Fig. 1 Example map of load equivalent (criterion, 21 days NOEC aquatic invertebrates)

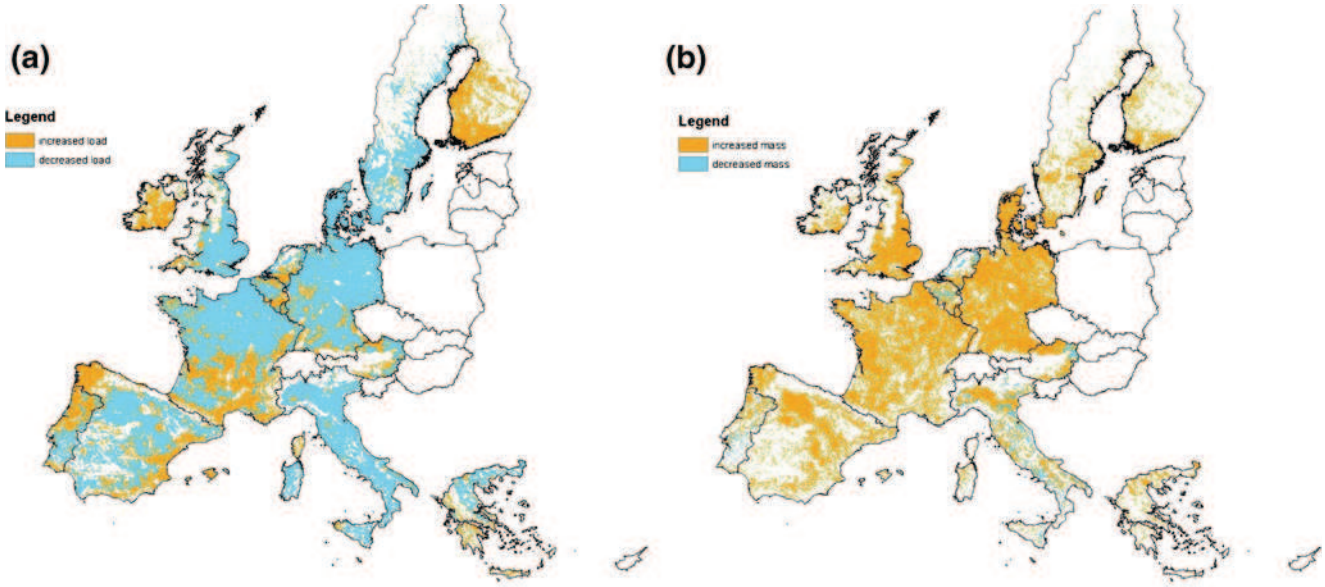


Fig. 2 Areas of increased/decreased pesticide load to streams (**a**) and mass in soil (**b**), expressed in omethoate equivalent and picoxystrobine equivalent, respectively

agricultural area (Oltrepo Pavese, in the southern part of the River Po basin in the Lombardia Region, Northern Italy), considering pesticide runoff and drift processes. The area includes seven river basins of tributaries of the River Po. Four herbicides (alachlor, terbuthylazine, metholachlor, and pendimethalin) have been selected in this case study as they are the most widely used pesticides applied on maize (the main crop in the area).

The distribution of predicted environmental concentrations (PECs) in surface water due to a single drift or runoff event for individual chemicals can be mapped at different resolution, using maps of environmental parameters (land use and crop distribution, application rate, river flows, etc.) at appropriate scale using well established procedures [41]. The four selected herbicides are applied in the same period (late April). Figure 3 shows the spatial distribution of the PECs produced by runoff after the first rain event after application. The basins of the seven rivers have been divided into sub-basins characterized by relatively homogeneous environmental parameters. The distribution of PECs reflects climate and landscape differences among basins and sub-basins (rain, crop distribution and density, slope, water flow, etc.), as well as different application rates and properties of the chemicals, and in this case, it can be considered reliable and also pointwise. Crop density and slope are the major driving forces responsible for herbicide

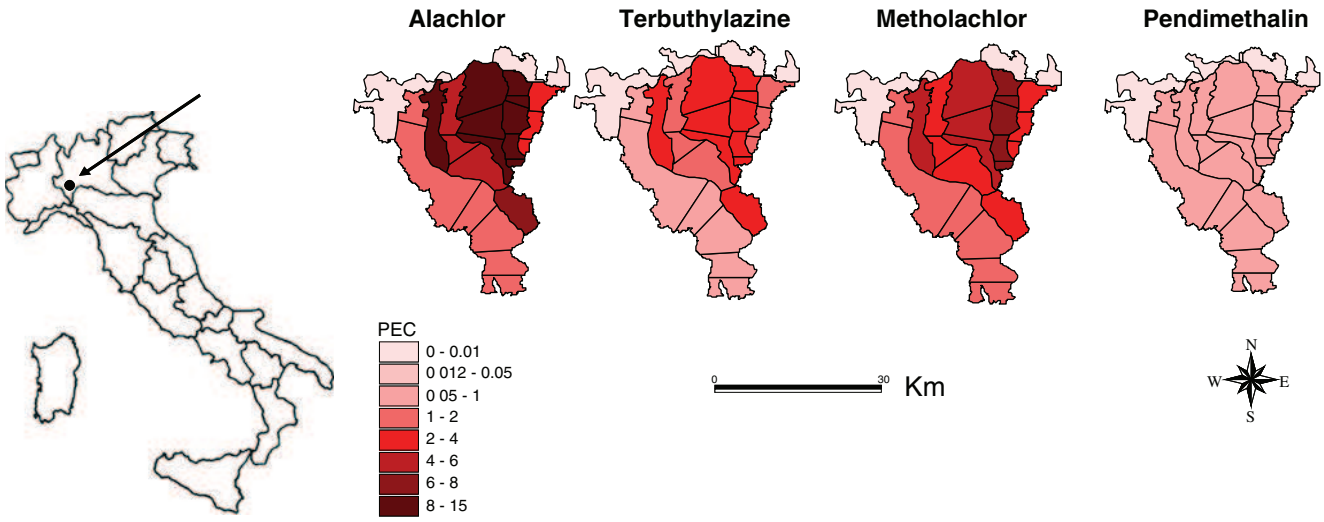


Fig. 3 Map of PECs (in μg/l) studying surface water for the selected herbicides due to runoff in the first rain event after application

runoff in surface water. The risk due to the mixture of the four selected herbicides was calculated with the CA approach. In Fig. 4, the risk for algae, calculated from 96 h EC_{50}, is shown as an example. The mixture toxicity map shows that in some basins, a potential risk for acute toxicity may occur. Considering the relative contribution of individual components of the mixture, alachlor and terbutylazine account for more than 80% of the total mixture toxicity.

The calculation conducted in the example can be applied to more complex mixtures, including possibly all pesticides used in a given agricultural area, and can be repeated for each significant emission event (rain events or drift corresponding to application). Moreover, the cumulative risk for all the components of the aquatic community (plants, invertebrates, and fish) can also be estimated by applying suitable risk indices for the biological community [15]. An example of application and validation of the procedure for the description of the time variability of PECs for individual chemicals is reported by Bonzini et al. [7]. A complete assessment of mixture composition from all the pesticides used in a pilot area during the whole productive season is reported by Verro et al. [55, 56].

As a *third example*, we consider the Dutch Environmental Indicator model for plant protection products, notably pesticides (denoted by the Dutch acronym NMI: [50, 52]) developed jointly by the Dutch National Institute for Public Health and the Environment (RIVM) and Alterra, Wageningen UR. This method is used to evaluate the impact of national pesticide reduction policies. We will show some results of NMI calculations for three test pesticides: chlorpyrifos, imidacloprid, and diazinon. The NMI was used to demonstrate the potential environmental impact of the separate substances and the substances combined for the year 1998 and to explore some of the visual possibilities of NMI maps (see Supporting information).

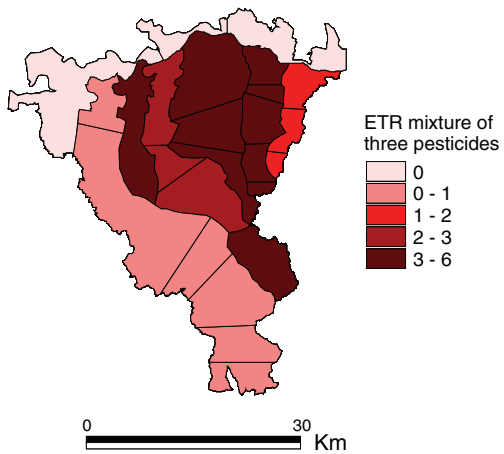


Fig. 4 Map of distribution of ETR for algae or TU, calculated for the mixture of the selected herbicides corresponding to the first rain event after application

The NMI estimates emission of pesticides to air, groundwater and surface water, potential acute and chronic effects on soil and water organisms, and potential contamination of drinking water by leaching to ground water. Based on crop data, the results of NMI calculations can be visualized for particular years in maps with 25-ha grid cells.

If one is interested in emission figures, for example for emission reduction programs, calculations can be focused on emissions rather than on potential ecological effects. Figure 5 shows the emissions in 1998 of atrazine, a herbicide that has since been banned in the Netherlands. The map shows where emission reduction measures are most urgent and will be most effective in terms of total national use. The map may be combined with a histogram showing the total number of grid cells in the country with emissions in certain categories. The histogram can be used to evaluate the amount of grid cells that are (still) above a certain threshold level, for example an emission of 1 kg atrazine per grid cell (Fig. 5).

Estimated environmental concentrations are also divided by environmental effect concentrations such as EC_{50} values. This yields TU values which are used as indicators of risk. For groundwater, estimated concentrations are normalized through the legislative standard for drinking water. In the NMI environment, this type of quotient, as well as the TU values for aquatic and terrestrial organisms, are jointly called environmental indicator units (EIU).

As in the previous examples, potential cumulative environmental impact can be evaluated with the CA assumption, i.e., by adding the EIU values. If one is interested in the overall environmental potential impact of chemical compounds, maps can be made as shown in Fig. 6. This map shows for the three test insecticides where in The Netherlands the EIU values for one of the indicators (aquatic, terrestrial, and drinking water) for a single substance exceeds 1. However, it also shows where the EIU for combinations of pesticides may exceed 1 (pink areas; it only occurred for combinations of imidacloprid and chlorpyrifos). These are additional risk areas that would not show up on a map for single pesticides.

The examples from the NMI presented here and in the Supporting information demonstrate some of the ways in which risks of pesticides can be displayed at a national scale. It is possible to visualize risks of single substances but also risks of several substances in one map. However, one is always limited by the number of categories that can be displayed in a single map. This is determined by the number of different colors that people can reasonably distinguish. So, the chosen display depends on the objective of the maps and the kind of information that needs to be

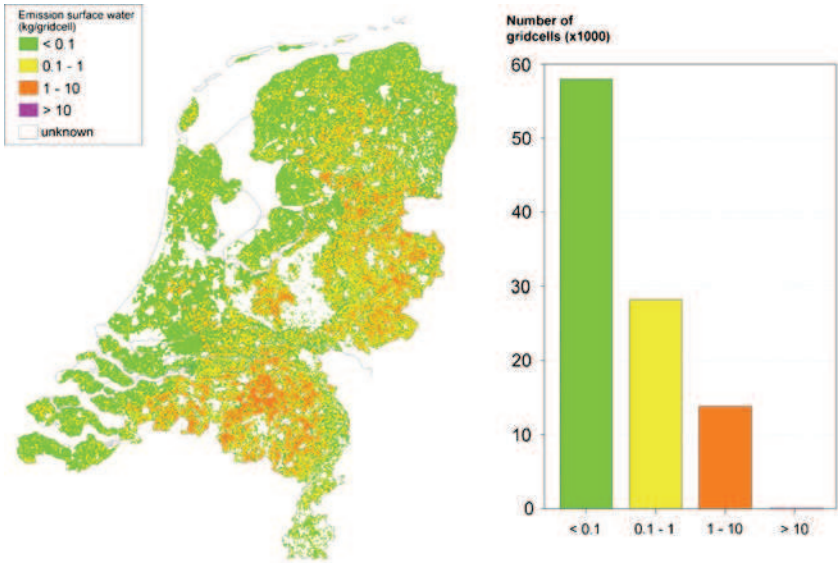


Fig. 5 Potential emission to surface water of the herbicide atrazin in 1998 calculated by the NMI

passed on to the users. Combination of emission or risk maps with other visualization methods such as graphs (Fig. 5) provides a powerful tool for communicating to decision makers, for instance when designing pesticide reduction schemes (e.g., [51]).

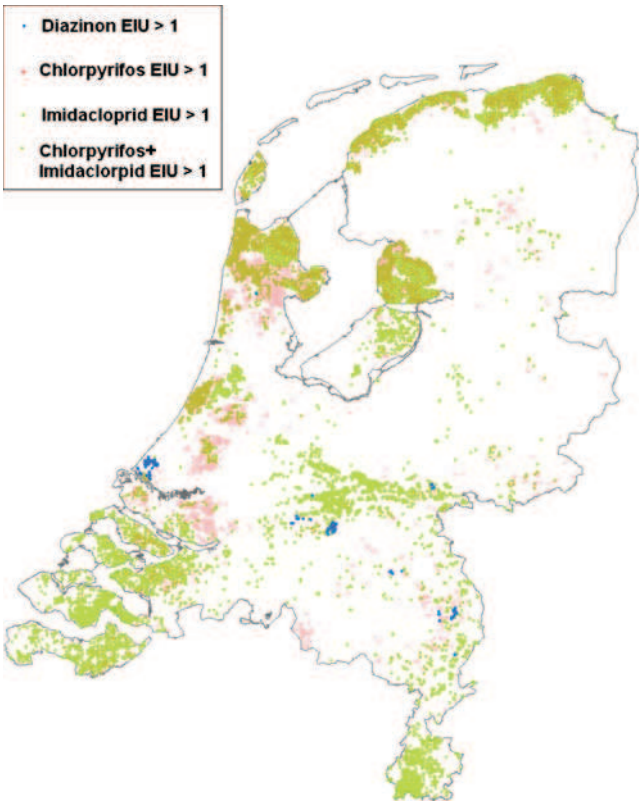


Fig. 6 Overall risk for three insecticides (chlorpyrifos, imidacloprid, and diazinon) in The Netherlands in 1998 assessed with the NMI. EIU are calculated by dividing PECs by threshold concentrations

3.2 Local Risk Mapping Based on Receptor-Oriented Modeling: An Example of Wildlife Exposure to Heavy Metals

This example presents a cumulative risk map that is a result of cumulative risk estimates of heavy metals for wildlife species using the receptor-oriented wildlife exposure model SpaCE [30, 43], grounding on the CA approach as discussed above. Such a receptor-oriented model simulates the exposure pattern of receptors based on their individual characteristics, e.g., the spatial foraging behavior, food preferences, and physiology (i.e., feeding rate). SpaCE consists of three main modules. The landscape module comprises the spatial input data for the model, i.e., species-specific habitat maps and maps of the contaminant concentrations in soil. The latter were made by inverse distance weighted (IDW) interpolation of point data, i.e., chemical soil concentrations measured in the study area. Further, the foraging module simulates the spatial foraging behavior. Movement algorithms simulate the receptor over a rasterized habitat map during the course of its life starting from a "nest" location. Finally, the exposure module simulates the contaminant flow in the food web. The internal contaminant concentrations in food items of the receptors (i.e., soil dwelling invertebrates, gastropods, and vegetation) are calculated using empirical relations, relating the internal contaminant concentrations to the concentrations in the soil. The lifetime average exposure concentration in food is calculated for each contaminant and every individual receptor obeying food web relations and depending on the local contaminant concentrations in the available food items encountered during the foraging. These PECs are compared with the predicted no effect concentrations in food (the threshold concentration) by computing HQs to

determine the risk from each contaminant. After this normalization, the risks are added up following the Concentration Addition approach (Eq. 1).

The cumulative risks for the individual receptors are plotted on a map as point estimates, where the nest location is used as the location to allocate the receptor. All risk point estimates (representing a population of multiple individuals) are then converted to a raster output. In this procedure, the same cell size of the soil contaminant concentration maps need to be assigned to the movement simulation model grid: In case more than one individual was modeled in one cell, the mean of the risk values of these individuals was assigned to the cell. To cover the whole area where the receptor species reside, the IDW interpolation method was applied for assigning risk values to cells in which no individuals were modeled.

In a case study in the Afferdensche and Deestsche Waarden floodplain in the Netherlands, 225,000 common shrew individuals were modeled in their suitable habitat. Assuming concentration addition, their cumulative risk to Cd, Ni, and Zn is shown in Fig. 7, through the coupling of a cartographic display with pie charts, effective for conveying proportion [19]. Color is an important visual attention guide and influences risk perception [59]. The level of risk (expressed as HQ) is visualized according to the risk hierarchy for color (i.e., red riskier than yellow, yellow riskier than green) found by Sattler et al. [42].

For comparison, risks from individual stressors are shown in Fig. 8. A receptor-oriented model, such as SpaCE, is an effective approach for addressing cumulative risk and can be used for risk mapping purposes. SpaCE estimates risk for receptors for substances that do not interact and in areas were these receptors forages (i.e., in suitable habitat). By simulating multiple individuals per nest location, the model can be used to estimate and map the variation around the risk as a result of foraging behavior.

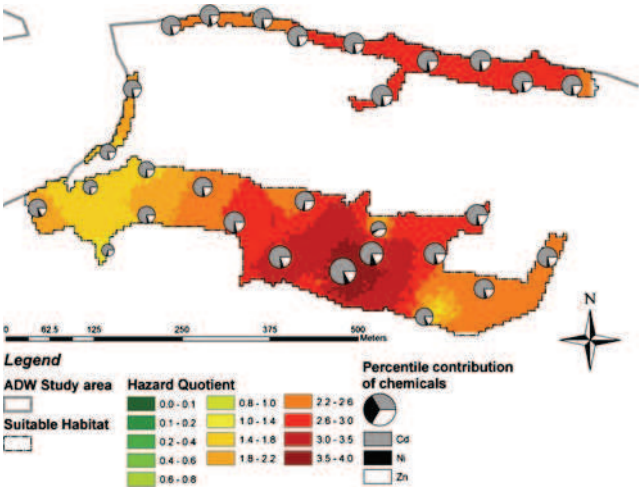


Fig. 7 Cumulative risk (HQ) of cadmium, nickel, and zinc, assuming concentration addition, to the common shrew (*Sorex araneus* L.) in a part of the Afferdensche en Deetsche Waarden study area

As the SpaCE model estimates the risk for mobile receptors, there is some uncertainty involved in the mapping of their risk; it involves a choice of allocating the risk onto a map. The predicted risk is actually a result of foraging behavior within the home range (i.e., represented by a rectangular area around the nest location). But, the home ranges overlap, making it difficult to map risk per home range onto a raster layer. Since the risk of an individual can be interpreted as the home range average risk and the nest location is always located at the center of the home range, it is considered justifiable to assign the risk of an individual to its nest location using the coordinates of the nest to plot the risk.

3.3 Data-Driven Risk Mapping: An Example on Air-Borne Benzene in Leipzig

The example presented here is based on measurements of ambient benzene concentrations at 191 dwellings in the city of Leipzig, Germany. Each site was sampled one time for a period of about 4 weeks (from January 2001 to April 2002).

To protect human health, in Germany, a limit for air pollution with benzene is set to 5 $\mu g/m^3$ [4]. Not exceeding this threshold, the median value of the measured benzene concentrations in Leipzig is 1.29 $\mu g/m^3$, similar to other German cities like Erfurt (median 1.62 $\mu g/m^3$) or Hamburg (median 1.13 $\mu g/m^3$) [45]. Srivastava et al. [46] observed in residential areas in the mega city of Delhi, India, a mean benzene concentration that is about ten times higher. The mean predicted benzene concentration for the city of Leipzig in December is 2.36 $\mu g/m^3$. Jo et al. [23] measured in residential areas in Daegu, South Korea during winter a geometric mean that is about three times higher than the predicted mean for Leipzig in December.

We observed a seasonal cycle in ambient benzene concentrations, with lower levels in summer than in winter, comparable to the results by Hansen and Palmgren [18]. Pekey and Arslanbas [33] found lower ambient benzene concentrations in summer than in winter in urban areas, offices, and schools. A comparable seasonal cycle of concentrations was observed by Schlink et al. [44] for indoor volatile organic compounds.

The measurements were processed using Bayesian inference. With a generalized linear regression, assuming a log-normal distribution of benzene concentrations, we take into account spatial correlation between the sampled sites [3, 48]. In this way, space is explicitly included in risk assessment. From a set of factors those with significant impact to the concentration were identified (Fig. 9), namely, (1) the factor "Land use" describing the type of land use,

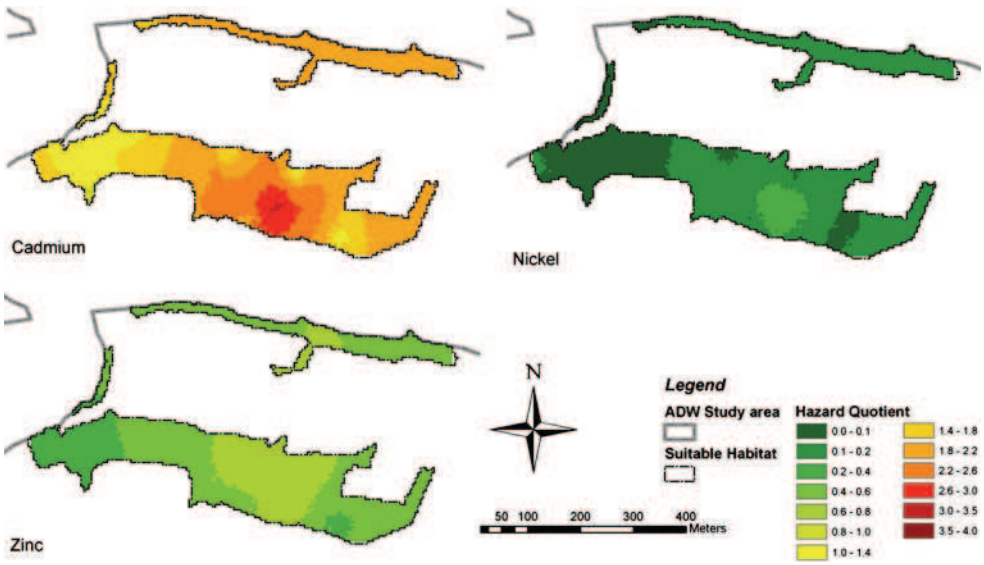


Fig. 8 Risk of cadmium, nickel, and zinc common shrew (*Sorex araneus* L.) in a part of the Afferdensche en Deetsche Waarden study area

i.e., green, water, agricultural, forest, residential, or industrial area; (2) "Traffic intensity" representing the traffic intensity; (3) "Bft" as a measure for the wind velocity; (4) "DistToCentr" representing the distance to the city center [km]; and (5) NE, SE, SW, and W for the frequency of winds coming from directions north-east, south-east, south west, and west, respectively. The values of the meteorological factors (Bft, NE, SE, SW, and W) are varying temporally; the values of the geographical factors (Others, TrafficNo50m, and DistToCentre) are varying spatially. The model was generally adjusted for the month of measurement. Model output was the monthly predicted benzene concentration at grid points with spacing 500 m for the city of Leipzig. Spatial interpolation yielded in geographical maps representing the continuous benzene concentration field (for December see Fig. 10).

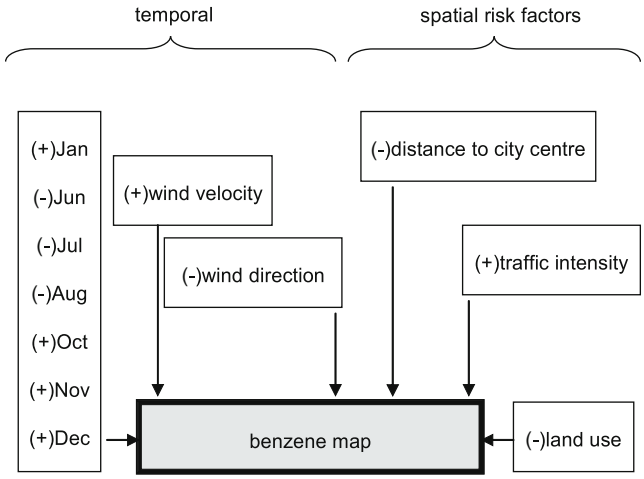


Fig. 9 Temporal and spatial risk factors with significant impact to the benzene concentration that were used for mapping (the sign indicates whether the factor has a positive or negative influence)

The statistical model provides a multiplicative decomposition of the cumulated risk into parts that are attributable to the individual factors (Fig. 10). In detail, we find that the higher Traffic or Bft, the higher benzene concentration. With increasing DistToCentre, the ambient benzene concentration is getting lower. The more often wind comes from NE, SE, SW, or W, the lower is the benzene concentration. The benzene pollution level varies with season and is significantly higher during winter (October to January) than in summer (May and June).

The spatial variation of benzene concentration in the maps for December is caused by the spatially varying impact factors. The high concentration around the city center corresponds to high traffic intensity, land use, and short distance to the city center. The influence of the traffic intensity is also reflected in the hot spots at the drive ups to the motorway in the north-east of Leipzig.

This statistical technique allows the consideration of scenarios of different severity: Median benzene concentration levels (Fig. 9, middle) reflect the ordinary case scenario. They are high in and around the city center, where traffic intensity is higher than in the peripheral regions of the city. In the north-east, there are three hot spots that are situated at drive ups to a motorway. In the worst-case scenario (Fig. 9, right hand side)—based on the 95th percentile of the predictions—the maximal benzene concentration is nearly twice as high as it is for the ordinary case scenario. If benzene concentrations come up to the values predicted in the worst-case scenario, an acute health risk at places around the concentration hot spots cannot be excluded. The location of concentration hot spots of the worst case agree with the ordinary case. From the best-case scenario (Fig. 9, left hand side), concentration hot spots are not identifiable. Spatial variation of the concentrations is rather limited; there is just background pollution all over

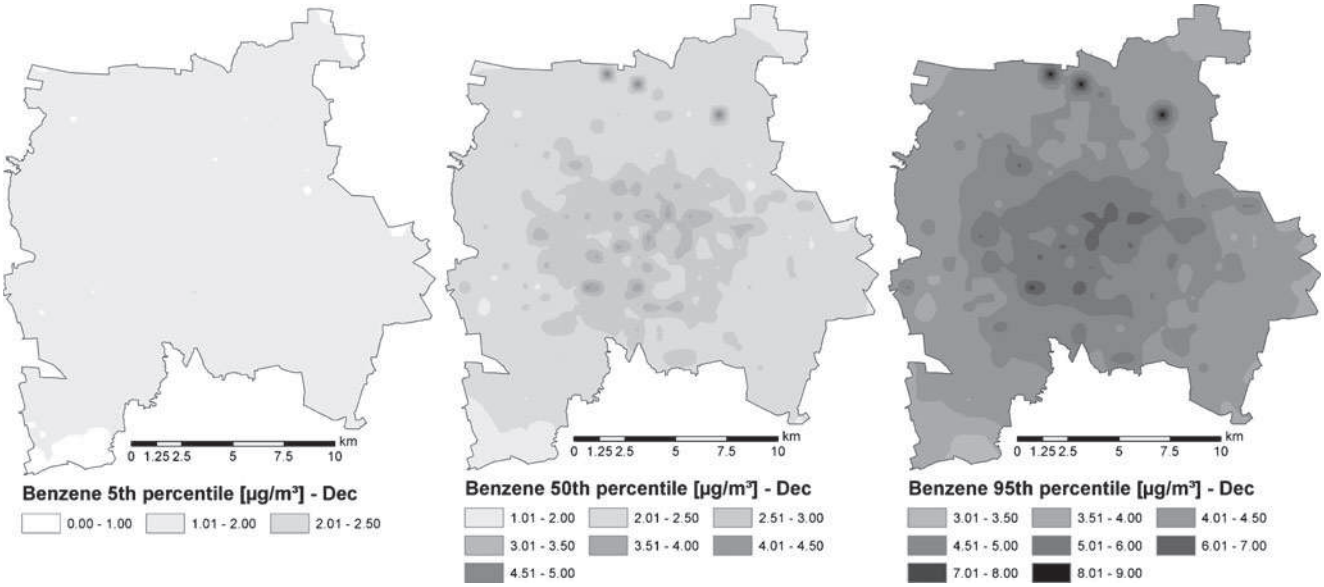


Fig. 10 Predicted benzene concentration field for the city of Leipzig, Germany in December. Best-case scenario, 5th percentile (*left hand side*); ordinary case scenario, 50th percentile (*middle*); worst-case scenario, 95th percentile (*right hand side*)

the town. The range of variation of the concentration is lowest for the best and highest for the worst-case scenario.

3.4 Classification Based on Known Risk Factors: Groundwater Vulnerability Mapping Using Self-organizing Maps

Our society is increasingly aware of the environmental status of aquifers since they provide one of the most important sources of potable water. The continuous emission of anthropogenic pollutants into the aquifer reduces water quality and may eventually threaten our drinking water supply.

The assessment of groundwater vulnerability is usually performed on the basis of vulnerability indicators reflecting individual factors affecting vulnerability, combined in order to obtain a comprehensive and synthetic characterization of the actual aquifer vulnerability. A widely used and well-known method is the DRASTIC index, developed by the US Environmental Protection Agency (EPA) as a standardized system [1]. The DRASTIC index is obtained by the weighted sum of seven hydrogeological properties, i.e., depth to water table (D), net recharge (R), aquifer media (A), soil media (S), topography (T), impact of the vadose zone (I), and hydraulic conductivity of the aquifer (C) as:

$$\text{DRASTIC index} = 5D + 4R + 3A + 2S + T + 5I + 3C \tag{3}$$

The seven hydrogeological variables are usually represented in maps, after transformation by rating each variable with values between 1 and 10. The variables increase with increasing vulnerability of the aquifer. The construction of maps for the seven variables requires extensive assessment, based on expert judgment and specific data described in detail in Aller [1].

We present an example for Camp de Tarragona, a hydrogeologic unit located in the southeast of Catalonia close to the Mediterranean Sea. It includes three counties, covers an area of 406 km^2, and has a dynamic economy with very important industrial, commercial, touristic, and agricultural activities. It includes two important cities, Tarragona and Reus, an airport, and an industrial harbor.

In the current study, the depth to water table was generated by kriging interpolation of the 315 piezometric data points available over region studied. Net recharge was calculated from the values of annual rainfall, land surface slopes, and soil permeability [34], which were accessible all in detail for whole area considered. Aquifer media information was obtained from complete geological maps, while soil media information was generated by kriging interpolation of only 123 infiltration capacity data points within Spain. Topography was obtained by processing a detailed digital terrain model with GIS. The impact of the vadose zone was calculated by linear combination of soil permeability and depth to water table [34]. Finally, the hydraulic conductivity parameter was inferred from the geological map by considering typical values of saturated hydraulic conductivity for the two dominant geological formations of rock type and grain size that exist in the Camp de Tarragona. The model of Eq. 3 is parametric, as it relies on weights for the different factors assigned a priori based on previous experience.

Here, we show how an automated classification method as the SOM procedure may be used when the a priori knowledge is not sufficient to apply parametric models to develop a vulnerability map of groundwater. Figure 11 illustrates the SOM-based vulnerability methodology applied to map aquifer vulnerability over the Camp of Tarragona area in Catalonia. At the left side of Fig. 11 are the trained SOM's C-planes for each variable, SOM U-matrix, and the Davies–Bouldin clustering [58] of the SOM units. The vulnerability map for the Camp of Tarragona based on SOM classification of the seven DRASTIC input parameters of the hydrogeological area is presented at the right side of Fig. 11. The optimal SOM configuration for this data corresponds to a hexagonal sheet map composed of 2,542 units with quantization and topological errors of 0.009 and 0.023, respectively.

Visual inspection of the SOM's C-planes reveals the presence of some correlations between variables, e.g., in the case of the parameters depth to water (D) and impact of the vadose zone (I). The distribution of high and low values over the C-planes are quite similar; the right area of the map has lower values than the left area for both parameters. Comparison of the net recharge (R) and the topography (T) reveals certain correlation in the high level values, especially at the lower part of the map. These correlations observed in the C-planes are evidence of the capability of SOM to find relations between variables because, as explained before, parameter I was calculated from D and S [34], where D is the most influential factor in the calculation. Also, parameter R is calculated from values of rainfall, soil's permeability (S), and soil's slopes (T) [34].

The U-matrix reveals the limits of cluster units indicated by the distance between values in adjacent neurons. In the U-matrix legend, red color indicates the highest Euclidean distance and thus represents cluster borders, while blue color indicates closer units representing compact areas. The Davies–Bouldin [58] index was used to select the optimal number of clusters based in an optimized K-means partition of SOM units. Five clusters were identified by the Davies–Bouldin index and labeled applying DRASTIC weights to data of each cluster center. Blue color was assigned to lower cluster value and red color to high cluster value, ranging from low vulnerability impact to high vulnerability impact in five distinctive classes.

4 Discussion and Conclusions

When considering mixtures of chemicals, a viable approach is to develop an explicit toxicity model. In practical

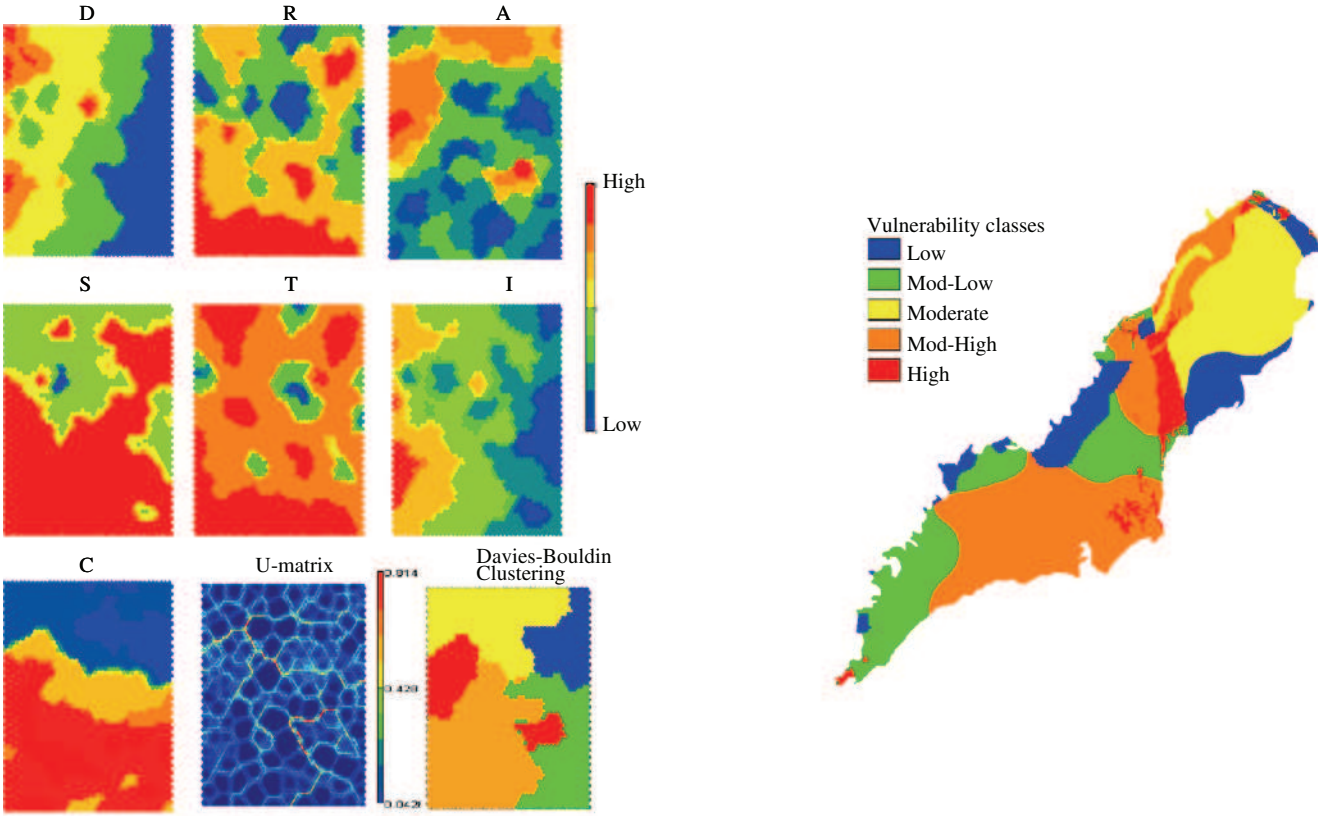


Fig. 11 SOM-based DRASTIC vulnerability index for the Camp de Tarragona area. (D depth to water, R net recharge, A aquifer media, S soil media, T topography, I impact of the vadose zone, C hydraulic conductivity of the aquifer)

applications, the CA approach provides a reasonable basis of assessment, although in principle more rigorous methods might be applied. Pesticides provide a representative example of substances often generating pollution, hence risks, through their combination in mixtures. Although conceptually straightforward and computationally simple, the mapping of cumulative risks under the CA assumption requires reliable mapping of emissions to the environment, prediction of concentrations for individual substances, and knowledge of the relative toxicity in order to compute the mixture toxic units. In some cases, and for specific chemicals, it has been shown that risk mapping can be done at a much more detailed level, by including foraging habits of organisms. In this case, models act as simulators to obtain an estimate of actual exposure, and not just potential exposure as in the cases presented on pesticides, where receptors are assumed to be uniformly distributed.

Methods involving the representation of receptor conditions or behavior, in principle, enable representing true risk but require information additional to the distribution of chemical concentrations, and are therefore typically more specialized and expensive. When no such detailed information is readily available, adopting reasonable safe-side assumptions on receptors may be preferable.

As a risk (contamination, exposure, and true risk) mostly cannot be measured at each location of a map, the observations, made at representative sites, can be interpolated in a reasonable way by data-driven approaches, which identify and involve the most important factors determining this risk, as in the example of benzene. In other cases non-parametric statistical techniques can be used as discussed, e.g., in Chung and Fabbri [8, 9]. Data-driven techniques have been widely applied in other contexts, for instance in the mapping of geo-environmental hazards (e.g., [36]).

A clustering technique can be applied to areas where no known impact occurred, but a risk is known to be caused by a series of factors; in such cases, areas with similar combinations of these factors can be identified, and this provides a first classification that can be used in decision support. The SOM method yielded continuous vulnerability classes despite the fact that no geographical coordinates were used in the training process. The classification of geo-referenced data by SOM and the labeling of the resulting macro-classes yielded vulnerability maps consistent with previous and well-accepted methodologies, such as DRASTIC. Additionally, SOM provides a good basis to select the most suitable set of variables for a specific area of concern since it effectively represents spatial regions of similar multivariable patterns that are identified and characterized by non-linear correlations between variables.

An important issue is the way in which cumulative risk maps are used: Sometimes, spatial distributions are meaningful in a statistical sense (i.e., they provide meaningful values for the mean, median, and percentiles of risk indicators), but the actual values assumed by the map at specific locations might be unreliable. For instance, the estimation of pesticide PECs presented at the European scale is not reliable due to the scale of assessment and the limitations in pesticide emission data, as thoroughly discussed by [35]: Maps represent only a statistically plausible distribution. Therefore, in risk communication, in such cases, it is suggested to avoid referring directly to the maps, but rather to their histograms (statistical parameters), using such addresses as "between 1992 and 2003, about 7% of European land has decreased toxicity of pesticides in soil" or "between 1992 and 2003, about 70% of European land has decreased toxicity of pesticides in stream ecosystems."

A cumulative map integrates the risk from multiple causes together, thus reducing the information to one map. Cumulative aspects of chemical risk arise when considering a mixture of different chemicals and other stressors, a single chemical with multiple sources, a combination of factors determining vulnerability, or a combination of the above circumstances. Cumulative risk maps usually convey one single content: They represent, on a qualitative, ordinal or quantitative scale, the level of risk at each point. Therefore, although largely conventional, they are rather unambiguous, easy to interpret and to convey to non-experts for use in decision support, compared with sets of separate maps for single causes of risk. Moreover, in some cases, they can be repeated at different times (see for example maps of herbicide risk at the local scale in Figs. 3 and 4), producing a picture of risk distribution in space and time. For example, if temporal variability of chemical emissions is known, the temporal variability of mixture composition can also be assessed, as described by Verro et al. [55]. On the other hand, from most of the examples presented in this paper, it appears that mapping cumulative risks is far from being an easy task: methods of cumulative mapping entail simplifications and assumptions that make the final maps usually less certain and robust than maps of individual risks. Cumulative maps should be regarded as practical products to convey information to decision makers, the general public, and other stakeholders. They are not always scientific products to be challenged with experiments and evaluation, but rather the results of conventional representations of which the realism should be always critically evaluated through expert judgment. However, at least at the regional or local scale, where the distribution of critical input data in space and time can be obtained with sufficient detail and reliability, cumulative risk maps may be a sound and practical representation of expected critical areas. When such data are not available in space to the desired level of reliability, if at least the frequency distribution of individual risks can be represented to some reliability,

computing cumulative maps may still be useful to produce a synthetic interpretation of complex interactions of individual risks.

Acknowledgments This paper contains considerations jointly developed by different partners of the NoMiracle project consortium. The individual case studies presented here are provided by single partners, to which the reader may refer for further details and for all scientific aspects not related to the specific topic of risk mapping: A. Pistocchi and P. Vizcaino for the European mapping of pesticides, S. Sala and M. Vighi for the case study on pesticides in Lombardy, J. Groenwold and J. Lahr for the one on pesticides in the Netherlands, M. Loos and A. Ragas for the case study on risks to individual organisms, U. Schlink and K. Strebel for the case of benzene in Leipzig, and M. Mujica and R. Rallo for the case on aquifer vulnerability mapping in Catalonia. J. Lahr coordinated the mapping exercises within the frame of NoMiracle Project work package 4.4, while A. Pistocchi coordinated the writing of the paper. The research was partly funded by the European Commission FP6 contract no. 003956 (NoMiracle IP: http://nomiracle.jrc.ec.europa.eu). Funding of Alterra, Wageningen UR, was also obtained from the Strategic research program "Sustainable spatial development of ecosystems, landscapes, seas and regions" financed by the Dutch Ministry of Agriculture, Nature Conservation and Food Quality (LNV).

References

1. Aller, L., Bennet, T., et al. (1987). DRASTIC, a standardized system for evaluating groundwater pollution potential using hydrogeologic setting. U.S. Environmental Protection Agency, EPA, Report 600/2-87-035; 1-455.
2. Bartels, C. J., & Van Beurden, A. U. C. J. (1998). Using geographic and cartographic principles for environmental assessment and risk mapping. *Journal of Hazardous Materials, 61*, 115–124.
3. Best, N., Richardson, S., & Elliott, P. (2003) Spatial epidemiology. Short Course, September 8–9.
4. BImSchV, 22 (2007) Verordnung zur Durchführung des Bundes-Immissionschutzgesetzes (Verordnung über Immissionswerte für Schadstoffe in der Luft). Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit.
5. Bliss, C. I. (1939). The toxicity of poisons applied jointly. *The Annals of Applied Biology, 26*, 585–615.
6. Boedeker, W., Drescher, K., Altenburger, R., Faust, M., & Grimme, L. H. (1993). Combined effects of toxicants: the need and soundness of assessment approaches in ecotoxicology. *Science of the Total Environment, 134*(2), 931–938.
7. Bonzini, S., Verro, R., Otto, S., Lazzaro, L., Finizio, A., Zanin, G., et al. (2006). Experimental validation of a GIS-based procedure for predicting pesticide exposure in surface water. *Environmental Science & Technology, 40*, 7561–7569.
8. Chung, C. F., & Fabbri, A. G. (1993). The representation of geoscience information for data integration. *Nonrenewable Resources, 2*(2), 122–139.
9. Chung, C. F., & Fabbri, A. G. (1999). Probabilistic prediction models for landslide hazard mapping. *Photogrammetric Engineering and Remote Sensing, 65*(12), 1389–1399.
10. De Lange, H. J., Sala S., Vighi M., & Faber J. H. (2010) Ecological vulnerability in risk assessment—A review and perspectives. *Science of the Total Environment* (in press)
11. de Zwart, D. (2005). Ecological effects of pesticide use in the Netherlands: modeled and observed effects in the field ditch. *Integrated Environmental Assessment and Management, 1*(2), 123–134.
12. Drescher, K., & Boedeker, W. (1995). Assessment of the combined effects of substances: the relationship between concentration addition and independent action. *Biometrics, 51*, 716–730.
13. Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives, 112*, 998–1006.
14. Faust, M., Altenburger, R., Backhaus, T., Blanck, H., Bödeker, W., Gramatica, P., et al. (2003). Joint algal toxicity of 16 dissimilarly acting chemicals is predictable by the concept of independent action. *Aquatic Toxicology, 63*, 43–63.
15. Finizio, A., Calliera, M., & Vighi, M. (2001). Rating systems for pesticide risk classification on different ecosystems. *Ecotoxicology and Environmental Safety, 49*, 262–274.
16. Finizio, A., Villa, S., & Vighi, M. (2005). Predicting pesticide mixtures load in surface waters from a given crop. *Agriculture, Ecosystems & Environment, 111*, 111–118.
17. Greco, W., Unkelbach, H. D., Pöch, G., Suhnel, J., Kundi, M., & Bodeker, W. (1992). Consensus on concepts and terminology for combined-action assessment: the Saariselkä agreement. *Archives of Complex Environmental Studies, 4*(3), 65–72.
18. Hansen, A. B., & Palmgren, F. (1996). VOC air pollutants in Copenhagen. *The Science of the Total Environment, 190*, 451–457.
19. Hollands, J. G., & Spence, I. (1998). Judging proportion with graphs: the summation model. *Applied Cognitive Psychology, 12*, 173–190.
20. Husdal, J. (2001) Can it be that dangerous? Issues in visualization of risk and vulnerability. http://husdal.typepad.com/blog/2001/10/can-it-really-b.html.
21. Ippolito, A., Sala, S., Faber, J. H., & Vighi, M. (2010) Application of vulnerability analysis; a case study of river basin. *Science of the Total Environment* (in press).
22. Jarosinska, D. (2009). Protecting human health and ecosystems—connecting novel research, practice and policy on multiple stressors Proc. NoMiracle/PHIME Conference "Multiple Stressors—Novel Methods for Integrated Risk Assessment" Aarhus, Denmark, 28th–30th September. http://nomiracle.jrc.ec.europa.eu/Documents/Conference_28-30_September_2009/Proceedings.pdf.
23. Jo, W. K., Lee, J. W., & Shin, D. C. (2004). Exposure to volatile organic compounds in residences adjacent to dyeing industrial complex. *International Archives of Occupational and Environmental Health, 77*(2), 113–120.
24. Junghans, M., Backhaus, T., Faust, M., Scholze, M., & Grimme, L. H. (2006). Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures. *Aquatic Toxicology, 76*, 93–110.
25. Kangas, J. A., Kohonen, T. K., & Laaksonen, J. T. (1990). Variants of self-organizing maps. *IEEE Transactions on Neural Networks, 1*, 93–99.
26. Kaski, S. (1997). Data exploration using Self-Organizing Maps. Dissertation for the degree of Doctor of Technology, Helsinki University of Technology, Espoo.
27. Kohonen, T. (1990). The self-organizing map. *Proc IEEE, 78*, 1464–1480.
28. Lahr, J., & Kooistra, L. (2009) Environmental risk mapping: state of the art and communication aspects. *Science of the Total Environment* (in press).
29. Lahr, J., & Kooistra, L. (2010). Environmental risk mapping of pollutants: state of the art and communication aspects. *Science of the Total Environment*. doi:10.1026/j.scitotenv.2009.10.045.
30. Loos, M., Ragas, A. M. J., Plasmeijer, M. J., & Hendriks, A. J. (2010) A receptor-oriented ecological exposure model for terrestrial vertebrates in an object-oriented programming platform. *Science of the Total Environment* (in press).
31. Moen, J. E. T., & Ale, B. J. M. (1998). Risk maps and communication. *Journal of Hazardous Materials, 61*, 271–278.

32. Nelson, P. (2000). Australia's national plan to combat pollution of the sea by oil and other noxious and hazardous substances—overview and current issues. *Spill Science & Technology Bulletin, 6*, 3–11.
33. Pekey, H., & Arslanbas, D. (2008). The relationship between indoor, outdoor and personal VOC concentrations in homes, offices and schools in the metropolitan region of Kocaeli, Turkey. *Water, Air, and Soil Pollution, 191*(1–4), 113–129.
34. Piscopo, G. (2001). Groundwater vulnerability map explanatory notes, Castlereagh Catchment. NSW Department of Land and Water Conservation, Australia.
35. Pistocchi, A., & Bidoglio, G. (2009) Is it presently possible to assess the spatial distribution of agricultural pesticides for continental Europe? A screening study based on available data.
36. Pistocchi, A., Luzi, L., & Napolitano, P. (2002). The use of predictive modeling techniques for optimal exploitation of spatial databases: a case study in landslide hazard mapping with expert-system-like methods. *Environmental Geology, 41*(7), 765–775.
37. Pistocchi, A., Vizcaino, P., & Hauck, M. (2010) A GIS model-based screening of potential contamination of soil and water by pyrethroids in Europe. Journal of Environmental Management. ISSN 0301-4797. doi:10.1016/j.jenvman.2009.05.020.
38. Pistocchi, A., Vizcaino, P., & Sarigiannis, D. Spatially explicit multimedia fate models for pollutants in Europe: state of the art and perspectives. Science of the Total Environment. doi:10.1016/j.scitotenv.2009.10.046
39. Plackett, R. L., & Hewlett, P. S. (1952). Quantal responses to mixtures of poisons. *Journal of the Royal Statistical Society. Series B, 14*, 141–163.
40. Price, P. S., Chaisson, C. F., Koontz, M., Wilkes, C., Ryan, B., Macintosh, D., et al. (2003). *Construction of a comprehensive chemical exposure framework using person-oriented modeling*. Annandale: The LifeLine Group. 129 pp.
41. Sala, S., & Vighi, M. (2007). GIS-based procedure for site-specific risk assessment of pesticides for aquatic ecosystems. *Ecotoxicology and Environmental Safety, 69*(1), 1–12.
42. Sattler, B., Lippy, B., & Jordan T. (1997) Hazard communication: a review of the science underpinning the art of communication for health and safety. US Department of Labor, Washington, DC. http://www.osha.gov/SLTC/hazardcommunications/hc2inf2.html. Accessed May 2009
43. Schipper, A. M., Loos, M., Ragas, A. M. J., Lopes, J. P. C., Nolte, B., Wijnhoven, S., et al. (2008). Modeling the influence of environmental heterogeneity on heavy metal exposure concentrations for terrestrial vertebrates in river floodplains. *Environmental Toxicology and Chemistry, 27*, 919–932.
44. Schlink, U., Rehwagen, M., Damm, M., Richter, M., Borte, M., & Herbarth, O. (2004). Seasonal cycle of indoor-VOCs: comparison of apartments and cities. *Atmospheric Environment, 38*(8), 1181–1190.
45. Schneider, P., Gebefugi, I., Richter, K., Wolke, G., Schnelle, J., Wichmann, H. E., et al. (2001). Indoor and outdoor BTX levels in German cities. *The Science of the Total Environment, 267*(1–3), 41–51.
46. Srivastava, A. (2005). Variability in VOC concentrations in an urban area of Delhi. *Environmental Monitoring and Assessment, 107*(1–3), 363–373.
47. Tait, N. G., Lerner, D. N., Smith, J. W. N., & Leharne, S. A. (2004). Prioritisation of abstraction boreholes at risk from chlorinated solvent contamination on the UK Permo-Triassic sandstone aquifer using a GIS. *The Science of the Total Environment, 319*, 77–98.
48. Thomas, A., Best, N., Lunn, D., Arnold, R., & Spiegelhalter, D. (2004). GeoBUGS user manual, version 1.2. Cambridge: Medical Research Council Biostatistics Unit; 2004. http://www.mrcbsu.cam.ac.uk/bugs/winbugs/geobugs.shtml.
49. Tortell, P. (1992). Coastal zone sensitivity mapping and its role in marine environmental management. *Marine Pollution Bulletin, 25*, 88–93.
50. Van der Linden, A. M. A., Luttik, R., Deneer, J. W., & Smidt, R. A. (2004). Dutch environmental indicator for plant protection products. Description of input data and calculation methods. Report no. 716601009/2004, RIVM/Alterra, Bilthoven/Wageningen, The Netherlands.
51. Van der Linden A. M. A., van Beelen, P., van den Berg, G. A., de Boer, M., van der Gaag, D. J., Groenwold, J. G., et al. (2006) Evaluation sustainable crop protection. Report nr. RIVM607016001, RIVM, Bilthoven, The Netherlands.
52. Van der Linden, A. M. A., Luttik, R., Groenwold, J. G., Kruijne en, R., & Merkelbach, R. C. M. (2008). Dutch Environmental Indicator for plant protection products, version 2. Input, calculation and aggregation procedures, Report nr. 607600002/2008, RIVM, Bilthoven, The Netherlands.
53. Van Leeuwen, C. J., & Hermens, J. L. M. (1995). *Risk assessment of chemicals: An introduction*. Dordrecht: Kluwer. 374 pp.
54. U.S. EPA (2006). Considerations for Developing Alternative Health Risk Assessment Approaches for Addressing Multiple Chemicals, Exposures and Effect (External Review Draft). U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-06/014A, 2006.
55. Verro, R., Finizio, A., Otto, S., & Vighi, M. (2009). Predicting pesticide environmental risk in intensive agricultural areas. II: Screening level risk assessment of complex mixtures in surface waters. *Environmental Science & Technology, 43*, 530–53.
56. Verro, R., Finizio, A., Otto, S., & Vighi, M. (2009). Predicting pesticide environmental risk in intensive agricultural areas. I: Screening level risk assessment of individual chemicals in surface waters. *Environmental Science & Technology, 43*, 522–529.
57. Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis, 3*, 11–126.
58. Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks, 11*, 586–600.
59. Wogalter, M. S., Conzola, V. C., & Smith-Jackson, T. L. (2002). Research-based guidelines for warning design and evaluation. *Applied Ergonomics, 33*, 219–230.
60. Wood, M., & Jelínek, R. (2007) Risk mapping in the new member states. A summary of general practices for mapping hazards, vulnerability and risk. Report no. EUR 22899 EN, Institute for the Protection and Security of the Citizen, Joint Research Centre, European Commission, Ispra, Italy, 26 pp.
61. Woodbury, P. B. (2003). DOs and DON'Ts of spatially explicit ecological risk assessment. *Environmental Toxicology and Chemistry, 22*, 977–982.
62. Worrall, F., & Besien, T. (2005). The vulnerability of groundwater to pesticide contamination estimated directly from observations of presence or absence in wells. *Journal of Hydrology, 303*, 92–107.

Supporting information

A document of supporting information is available along with the paper, containing additional material for the illustration of the case studies presented here.

A.2 Paper on SOM-based vulnerability in the Mediterranean region to be submitted to Environment International

Groundwater Vulnerability Assessment using Self-Organizing Maps

Marelys Mujica[a], Robert Rallo[b] and Francesc Giralt[a,*]

[a]*Departament d'Enginyeria Quimica, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain.*
[b]*Departament d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain.*

ABSTRACT

Intrinsic vulnerability maps identify geographical areas that are "vulnerable" or sensible to pollution based on intrinsic hydrogeological characteristics of aquifers and overlying media, independently of the pollutant characteristics. These vulnerable areas can be made more pollutant specific by also considering some characteristics of the overlying physical media related to potential pollutant sources (e.g., land uses) over the area of study. A new vulnerability index, both intrinsic and specific, was developed for groundwater vulnerability assessment by using Self-Organizing maps (SOM) without requiring previous expert's knowledge ratings of numerical variables as in the DRASTIC approach. The clustering capabilities of SOM facilitates dealing, in an integrated manner, with missing data, different sources and types of information, spatial interpolation, probabilistic analysis, variable selection and with the vulnerability assessment as well. The SOM vulnerability index was made independent of the scale of the geographical region considered by appropriate standardization and discretization of variables. The proposed scale-independent methodology was successfully tested at the local scale of the Camp de Tarragona hydrogeological unit and at the larger regional scale of Catalonia.

***Keywords*:** Groundwater vulnerability; Intrinsic Vulnerability index; Specific Vulnerability index; Self-organizing maps

---

Abbreviations: A, aquifer media; D, depth to water table; R, net recharge; S, soil's media; T, topography; I, impact to vadose zone; C, hydraulic conductivity of the aquifer; SOM, self-organizing map.

* Corresponding author at: Departament d'Enginyeria Quimica, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalunya, Spain. Tel.: +34 977 558 549; fax: +34 977 559 621.
E-mail address: francesc.giralt@urv.cat

1. Introduction

Groundwater is a major resource of potable water for human consumption, and stakeholders are worldwide concerned about its quality and degree of vulnerability. Industrialized and developed countries generate high loads of potential contaminants that, together with stressors, can negatively impact on the quality of their groundwater resources. Pollutants may reach groundwater usually by leaching through the ground and occasionally by direct contamination of wells. The fate of chemicals once they reach the soil depends on their physical and chemical properties, the degradation patterns and soil intrinsic characteristics (Mackay et al., 1996). The risk for groundwater contamination also depends on the incidental occurrence and the consequences of contamination events (Lindström, 2005). Worrall et al. (2002) studied the interrelationships between the chemical properties of contaminants (e.g., solubility in water) and site properties (e.g., soil and aquifer type, and land use) using analysis of variance (ANOVA). Their results provide statistical evidence that both molecular and site properties methods together (intrinsic vulnerability) have the potential to facilitate the understanding of groundwater pollution, as has been the case in pesticide contamination. The risk of groundwater pollution has also been evaluated directly from field measurements of contamination (Worrall and Kolpin, 2003; Worrall and Besien, 2005).

Intrinsic vulnerability maps define geographical areas that are "vulnerable" or sensible to pollution, independently of the pollutant characteristics. On the other hand, specific vulnerability assessment also considers stressors, such as the land uses, that are related to specific contamination scenarios in the analysis (Martínez-Bastida et al., 2010). The availability of groundwater intrinsic vulnerability maps should help decision-makers (governments, national environmental agencies) in land planning for agricultural, industrial or urban use (Villa and McLeod, 2002; Passuello et al., 2012). Also, remediation plans for specific pollutants can be supported by using regional specific vulnerability maps to screening the quality of groundwater resources (Aller et al., 1987; Perles Roselló et al., 2009; Vías et al., 2010; Zabeo et al., 2011).

Monitoring groundwater quality is a common practice to address aquifers' health. In many countries geo-referenced water quality data are publicly available and include information on the concentration of heavy metals, nitrates, nitrites and sometimes pesticides. The monitoring strategies are usually established as a function of the final use of groundwater resources, mostly human consumption, and are implemented by measurements at fixed locations. Quality indicators include the legislative thresholds for pollutants in potable water which are established by local, regional and supra-national environmental administrations.

Risk of contamination is the likelihood or probability that a contaminant actually present in the groundwater causes adverse effects at established end-points. Lahr and Kooistra (2009) reviewed the most important types of risk maps, including concentration maps, exposure maps and vulnerability maps. Mapping the groundwater quality information with concentration maps provides a picture of the actual (and/or past) state of the aquifers due to the actual (and/or past) discharge or presence of pollutants sources in

each area. Regions with high pollutant's concentration can point out high vulnerable zones and/or areas with an intensive load of pollutants. On the other hand, areas with low or undetectable concentrations can indicate either that the zone has low vulnerability or that the load of pollutant over that specific area is low or null. Concentrations maps can be used to evaluate vulnerability maps following these rules: high pollutants' concentrations *must* occur in high vulnerability areas and consistent vulnerability maps should minimize the occurrence of high pollutant concentrations in "low" vulnerability areas.

Vulnerability maps have been used worldwide to assess the risk of groundwater contamination (Almasri, 2008; Andreo et al., 2005; Martínez-Bastida et al., 2010; Martínez-Santos et al., 2008; Masetti et al., 2009; Neukum et al., 2008; Rahman, 2008; Sinan and Razack, 2009). Several models for vulnerability assessment have been reported in the literature: Overlay and algebraic methods like DRASTIC (Aller et al., 1987; Burchart et al., 2006; Draoui et al., 2008; Goldscheider, 2005; Moore, 1990), statistical approaches (Panagopoulos et al., 2006; Worrall and Besien, 2005; Worrall and Kolpin, 2003; Nolan et al., 1997), fuzzy methods (Dixon, 2005a; Dixon, 2005b; Gemitzi et al., 2006; Mao et al., 2006; Mazari Hiriart et al., 2003; Uricchio et al., 2004), and process based simulation models (Lindström, 2005; Martínez-Santos et al., 2008; Nolan and Hitt, 2006; Popescu et al., 2008). First principle models should always be preferred but are difficult to apply to real conditions. On the other hand, overlay and algebraic methods have been developed to overcome the limitation of the large amount of monitoring data that statistical methods require, but at the cost of not providing information on the uncertainty of the vulnerability predictions.

Many previous studies have explored the application of geostatistics techniques to generate smooth concentrations maps from field data by using different spatial interpolation methods (Tutmez and Hatipoglu, 2010; Kazemi and Hosseini, 2011). Hu et al. (2005) applied geostatistics techniques (kriging) to analyze the spatial variability of NO_3concentrations in groundwater. Also, different types of artificial neural networks have been applied in spatial interpolation methods. Zare et al. (2011) examined the advantages of the multilayer perceptron network for nitrate concentration mapping compared to linear regression methods. Self-Organizing Maps (SOM) (Kohonen, 1990; Kohonen, 2001; Vesanto and Alhoniemi, 2000) have been successfully applied in Environmental Risk Assessment for nitrate pollution in groundwater (Rallo, 2007). This author specifically studied the interpolating and recovering of missing data capabilities of SOM for groundwater risk assessment of nitrate contamination. SOM were used to generate smooth concentrations maps by mimicking the kriging and co-kriging techniques from point source data. SOM have also been applied in many hydrogeological applications (Céréghino and Park, 2009; Kalteh et al., 2008; Peeters et al. 2007; Sánchez-Martos et al., 2002).

The aim of the current study is to apply the SOM algorithm to classify intrinsic media variables, complemented or not by additional specific land information, to address groundwater vulnerability assessment in an integrated manner. The classical DRASTIC index analysis (Aller et al., 1987) has also been performed to compare the proposed SOM-based vulnerability index with a standard and worldwide used methodology (Babiker et al.,

2005). The new index has been defined and applied to both intrinsic and specific vulnerability analyses at two different spatial resolutions to evaluate the robustness of the proposed SOM methodology to changes in scale. These two different scenarios are the local area corresponding to the Catalan hydrogeological unit of the "Camp de Tarragona" and the larger region of Catalonia. The highest spatial resolution of the local scale has been first used to demonstrate the capabilities of the SOM to generate intrinsic vulnerability maps compared to DRASTIC and to provide the basis for a new and better approach for both intrinsic and/or specific vulnerability assessments. Second, the methodology has been applied to the regional scale of Catalonia for further evaluation by exploring the interpolation capabilities of SOM compared to kriging and its performance under up-scaling conditions. Both nitrates concentration and cumulative maps (i.e., concentration maps of exceeding legislative thresholds of different contaminants) are used for the evaluation of the proposed SOM vulnerability methodology.

2. Materials and methods

2.1 Area of study and data

The study is focused on Catalonia that is located in the north-east of Spain (Fig.1) with an area of 32,107 km^2. The total population is approximately 7,360,000 (population census of 2008), with near the 67% concentrated in the metropolitan area of Barcelona. Catalonia has a very diverse orography, having a long coastline (547 km long), extensive mountain chains mirroring the coastline, mountains peaks reaching 3000 meters in altitude in the Pyrenees, inland depressions and the sedimentary delta of the Ebre river. The climate is closely related to this orography. Winters are usually mild (average temperature of 6-7 ºC) and summers are hot and dry (average temperature of 24 ºC), with temperatures varying considerably between the coastline, the inland plains and the Pyrenees.

The regional scale of the current study covers the whole of Catalonia. The Camp de Tarragona, one of the 49 independent hydrogeological units of this region, was selected for the local scale study (Fig. 1). Geophysical characteristics were obtained from the Catalan and Spanish Governments. Hydrogeological maps were provided by the Catalan government, Departament de Medi Ambient i Habitatge and the Institut Cartografic de Catalunya (ICC), and the Spanish government, Instituto Geológico y Minero de España (IGME) among others. Table 1lists the sources, type of data and resolution of all available hydrogeological data used in the current study. All available geo-referenced intrinsic media characteristics, like geological profiles, digital terrain model, annual rainfall maps, and land use were processed by Geographical Information System (GIS) software to generate raster layers with a resolution 200 by 200 meters for the regional scale and 50 by 50 meters for the local scale. The GIS MiraMon (Pons, 2006) was used for visualization and mapping. Several informative maps are provided as Supporting Information. The spatial coordinates of the regional area of Catalonia are defined by the UTM-31N-UB/ICC reference system (UTM: Universal Transverse Mercator coordinate system, UB: University of Barcelona, and ICC: Cartographic Institut of Catalonia) with an areal spam of 258000 to 526600 meters in X-coordinate, and 4485000 to 4752000 meters in Y-

coordinate. The raster maps generated a total of 1792905 cells and 802739 active cells over Catalonia.

The local scale of the Camp de Tarragona, a hydrogeological unit with an area of 406 km^2, includes three counties (Tarragones, Alt Camp and Baix Camp). The area has a very dynamic economy with very important industrial, tourism related and agricultural activities in the context of Catalonia and the south of Europe. It includes two important cities, Tarragona and Reus, an airport and an industrial harbor. The total number of inhabitants is approximately 400,000. The dominant economic activities are industrial (41.6%) and services (43.3%), with real estate and construction (12%) and agriculture (3.1%) as the less important economic sectors. The spatial coordinates of the local area are defined by the UTM-31N-UB/ICC reference system with an areal spam of 320500 to 367800 meters in X-coordinate, and 4537500 to 4586000 meters in Y-coordinate. The spatial resolution of all raster maps at the local scale developed in this study was 50 by 50 meters, generating a total of 917620 cells and 261479 active cells over the region in study.

Groundwater data was provided by the Catalan Water Agency (ACA) for years 2002. This data set contains pollutant concentrations measured in 765 groundwater quality control points unevenly distributed over Catalonia (Fig. 2), of which 110 correspond to the Camp de Tarragona. Data included heavy metals (Al, Sb, Ar, Cd, Cr, Mo, Ni, Fe, Mn, Se, Ba, Cu, Pb, Zn, and Mg), nitrate, nitrite and sulfates as presented in Table 2. The groundwater quality control network defined by Catalan Water Agency is subdivided into sub-networks comprised of wells from the same aquifer or territorial area delimited by hydrogeological, geographical or other criteria. The basic network is formed by locations where measurements of physico-chemical composition of groundwater derived from water-environment interactions and diffuse pollution phenomena (not attributable to specific sources) were carried out. Also a specific nitrate-vulnerable area network was established by the Catalan government. It recorded measures of the majority of anions and cations, nitrogenous compounds (NO_3, NO_2, and NH_4) and metals in feed for animal consumption (Fe, Mn, Co, Zn, Se and Cu).

The conjunction of pollutant exposure with exceeding legislative threshold values has permitted the assessment of cumulative risk for groundwater pollution and the generation of probability maps. Cumulative risk maps were generated by addition of probability maps of those pollutants that exceed legal threshold. Geostatistics was used to create continuous surfaces from spatially distributed sample data. A theoretical model was fitted to sample the variogram for each variable. Ordinary and Universal kriging were performed to generate concentration maps. The cumulative effects of pollutants can be studied by identifying elevated risk areas in cumulative exposure maps where high concentrations of different contaminants in water converge. Boolean concentration maps were created by assigning a value of one if the value exceeded legal threshold in the concentration maps and zero otherwise. Fig. 3 depicts the smooth cumulative exposure map for the Camp de Tarragona, which was generated by superposition of Boolean concentrations maps. Twelve pollutants exceeded in at least one monitoring station the legal threshold for year 2002. As

expected, the highest concentrations of pollutants are found in the vicinity of the important industrial area and of the harbor, both located next to the city of Tarragona.

2.2 Vulnerability assessment and the DRASTIC index

Concepts for vulnerability, exposure and risk are stated in the following definitions in agreement with the conclusions and recommendations of the International Conference on Vulnerability of Soil and Groundwater to Pollutants (Duijvenbooden and Waegeningh 1987) and the work of Lahr and Kooistra (2010). Groundwater vulnerability uses the presence and geographical distribution of sensitive receptors of stress to map more and less vulnerable areas. Intrinsic vulnerability can be defined as the assessment of vulnerability areas based on intrinsic hydrogeological characteristics of aquifers and overlying media. On the other hand, specific vulnerability is defined as the assessment of vulnerability areas based on the above intrinsic information together with specific information of stressors, like the land uses, that can be related to sources of pollutants over the area of study.

The assessment of groundwater vulnerability is usually performed on the basis of vulnerability indicators. Different techniques have been implemented to generate these indicators by combining several vulnerability-related variables so that a comprehensive and synthetic characterization of the actual aquifer vulnerability could be obtained. The two most popular techniques are the overlay and the index methods, followed by statistical methods. The overlay and index methods are simple mathematical models, consisting in algebraic operations of hydrogeological parameters. Several models such as DRASTIC (Aller et al., 1987), SINTACS (Civita, 1994), SEEPAGE (Moore, 1990), and PI (Goldscheider, 2005) have been developed. Fuzzy methods estimate weights and rates for index methods using Fuzzy rules (Dixon, 2005a,b; Gemitzi et al., 2006; Mao et al., 2006; Mazari Hiriart et al., 2006; Uricchio et al., 2004). Statistical methods estimate groundwater vulnerability by statistical analysis of point pollution data by determining the statistical dependence between observed contamination, observed environmental conditions and observed land uses that are potential sources of contamination (Panagopoulos et al., 2006; Worrall and Besien, 2005; Worrall and Kolpin, 2003). Process-based simulation models use the governing equations for water flow and solute transport. The focus is on computing travel times or concentrations of a contaminant in the unsaturated and groundwater zones (Lindström, 2005).

The DRASTIC Index was developed by the EPA (Environmental Protection Agency of United States of America) as a standardized system for evaluating groundwater vulnerability to pollution (Aller et al., 1987; Babiker et al., 2005). It considers seven hydro-geological properties:

Depth to water table (D): is the distance from the ground to the water table. The shallower the depth the more vulnerable will be the aquifer to pollution.

Net recharge (R): measures the total quantity of water per unit area which reaches the water table. Recharge is the main vehicle for leaching and transport of any contaminant to the water table. Vulnerability is thus enhanced by high recharge rates.

Aquifer media (A): refers to the properties of the rock that serves as aquifer. Lithology and grain size is determinant for the transport of pollutants within the aquifer. The property of a rock to be pervaded by a fluid is called permeability. Also porosity plays an important role in vulnerability assessment. The higher the permeability and porosity in the aquifer, the higher will be its vulnerability to contamination.

Soil media (S): provides the characteristics of the upper weathered zone of the earth, the first 1.5 meters from the ground surface. The content of clay and organic matter are relevant in controlling the pollutant infiltration to aquifers. In general, the presence of clay and small grain size reduces the vulnerability of groundwater to pollution.

Topography (T): refers to the slope of the land surface. Vulnerability to contamination is reduced as the slopes increases due the increment in the runoff capacity of the media.

Impact of the vadose zone (I): provides information on the unsaturated zone above the water table. The texture of the vadose zone determines the travel time of pollutants through this zone.

Hydraulic conductivity of the aquifer (C): refers to the rate at which water flows horizontally through an aquifer. Vulnerability is increased as hydraulic conductivity increases.

The DRASTIC vulnerability mapping involves overlaying the seven hydrogeological properties. Each DRASTIC feature is assigned a weight relative to each other in an increasing range of importance, from 1 to 5, based on expert criteria as described in the following equation,

$$\text{DRASTIC Index} = 5D + 4R + 3A + 2S + T + 3I + 5C \quad (1)$$

Ratings assign values between 1 and 10 to each DRASTIC variable, in an increasing order of impact to vulnerability. This step requires a laborious pre-processing task for each variable, based on expert criteria and knowledge of the process under study. Piscopo (2001) presented a general methodology to generate DRASTIC features from common hydrogeological data. Many authors have applied this methodology worldwide: A shallow aquifer in India (Raman, 2008); the Senirkent-Uluborlu basin in Turkey (Senar et al., 2009) and urban watersheds in Mexico (Bojórquez-Tapia et al., 2009). There are many other contributions to generating site specific weights taking into account parameters such as land use or agricultural activities (e.g., Secunda et al., 1998; Umar et al., 2009).

Different vulnerability class labeling approaches have been proposed for the DRASTIC index. Aller et al. (1987) adopted three vulnerability categories of low, medium and high by using two percentile cuts of 48% and 68% in the vulnerability index distribution. Draoui et al. (2008) accounted for more vulnerability differentiation with the five classes to very low, low, low moderate, moderate and high, with four equi-espaced percentile cuts of 20%. In Ahmed (2009) labeling approach, only the four vulnerability classes low, moderate, high and very high, with percentile cuts of 50%, 65% and 75%, were defined. The final use of the groundwater vulnerability map determines which type of labeling is

more appropriate (Lahr and Kooistra, 2009). Policy-makers, environmental agencies and local governments are interested in a global evaluation of groundwater vulnerability. At that policy level, vulnerability labeling using three major categories (low, moderate and high) is preferred for its simplicity and has been also been the case in the current study.

2.3 Self-organizing maps

The self-organizing map (SOM) (Kohonen, 1990) is an unsupervised classifier for clustering and visualization of high dimensional data. A SOM consists of a number of units placed in a 2 or 3 dimensional grid. Units are connected to each other by a neighborhood relationship that governs the overall map structure forming either a rectangular or a hexagonal lattice. The SOM algorithm is based on competitive learning (Kangas et al., 1990; Kaski, 1997; Kohonen, 1990) where units gradually become sensitive to different input categories in a domain of the input space. Each unit is represented by a prototype vector $m_i=[m_{i1}, m_{i2}, ..., m_{in}]$ where n is the dimension of the input space. Prototype vectors are initialized using randomly selected input data. The SOM development begins with the random selection of an input data point *P* and computes the distances between *P* and the prototype vector of every SOM unit; the prototype vector of the unit closest to *P* (usually referred as Best Matching Unit or BMU) is adapted based on the distance to the BMU unit in the initial geometry. This preserves the topology or structure of the map. The SOM update procedure displaces the BMU and its neighboring units towards the input vector *P*. This updating process preserves the topology (structure) of the map. The update rule for the prototype vector of unit *i* is given by,

$$m_i = m_i+\alpha(t)h_{ci}(t)(P-m_i) \quad (2)$$

where m_i is the BMU, $h_{ci}(t)$ is a Gaussian neighborhood function, and $\alpha(t)$ a monotonically decreasing learning rate. The basic idea in the SOM learning process is that, for each sample input vector *P* presented to the map for classification, the elements of the prototype vector of the corresponding BMU (or class) as well as those of the neighborhood nodes are updated to incorporate the new information, i.e., the prototype vector of the BMU and those of its neighboring units become closer to the input vector. In the early stages of training, the radius that defines the size of the neighborhood is large; and most of the SOM units strongly belong to any node's neighborhood. This creates an initial good global ordering of the SOM. As the training progresses, the radius is reduced to yield good local ordering as well.

Distances between map units reflect their topological relationships and are defined by the difference vectors between the prototype vectors representing the input samples respectively clustered in each one of them. The adjacent neurons or units of any neuron m_i define its neighborhood $N_i(t)$. Its value is generally defined as a decreasing function of the number of iterations or times (epochs) that the input data are presented to the map for training. On the other hand, the learning rate in equation (2) quantifies the fraction of the learning needed by each neuron at each training iteration. A complete description of the SOM and details of its implementation are provided in Kohonen (1995).

The SOM is both a powerful classification and visualization tool that provides several good alternatives for visualization purposes. The unified distance matrix, or U-matrix, is perhaps the most used method to display SOMs. In addition, each SOM node can be easily related to the input data space by using the component planes. Straightforward correlations and relationships in the input data can be found by comparing several component planes (c-planes) at the same time (Kaski, 1997; Kaski et al., 1999; Laine, 2003). Since SOM represents the similarity clustering of multivariate attributes, the visual representation becomes more accessible and easy to use for exploratory analysis. This kind of spatial clustering facilitates the exploratory analysis of data with the purpose of identifying cause-effect relationships or correlations in exposure related problems when used in conjunction with environmental, transport and geophysical data.

Usually, the quality of a SOM is evaluated by the mapping precision and its topology preservation capabilities. The *mapping precision* describes how accurately the neurons 'respond' to the given data set. It is represented by the quantization error, *qe*,

$$qe = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i + m_c\right\| \tag{3}$$

Where *x* is a sample vector of the input space and *m* is a reference vector representing any given unit within the map. Lower quantization errors are associated to more accurate neuron responses in the SOM units. The topology preservation error, *te*, indicates how well the SOM preserves the topology of the data set,

$$te = \frac{1}{N}\sum_{i=1}^{N} u(x_i) \tag{4}$$

where $u(x_i)$ is 1 if the first and second BMUs or winning neuron (unit) are not next to each other, i.e., neighbors. Thus lower topographic errors imply better SOM topology preservation.

The SOM-based approach has been developed and assessed first at the local scale of the Camp de Tarragona. At this high resolution it is easier to demonstrate the capabilities of SOM to generate vulnerability maps (i) by mimicking the DRASTIC methodology in a sounder vulnerability class identification approach and (ii) by using the definition of a new vulnerability index herein called SOM-based approach. The regional scale information has been used to study the interpolation capabilities of SOM and to evaluate the behavior of the new vulnerability index during the up-scaling process.

3. Local scale vulnerability assessment

3.1 DRASTIC

The DRASTIC vulnerability index was calculated for the Camp de Tarragona at each grid cell following Equation (1), i.e., by weighted addition of variable layers. As each

DRASTIC feature is rated in a 1 to 10 scale, the resulting index will always be within the range 23 to 230. Detailed DRASTIC features and methodology for the Camp de Tarragona are presented in the Supporting Information. Fig.4 shows the DRASTIC intrinsic vulnerability map for the Camp de Tarragona obtained for the three categories of low, moderate and high proposed by Aller et al. (1987).

The evaluation of the map depicted in Fig. 4 or that of any intrinsic vulnerability map is not a simple task because there is not a direct measure of the vulnerability for a given area or region. Many authors have used nitrate concentrations in groundwater as an indication of the vulnerability of an area (Stitger et al., 2006; Mishima et al., 2011; Tilahun and Merkel, 2010). In the current study vulnerability maps have been evaluated by using the cumulative effect of different pollutants and/or the concentrations of nitrates exceeding legislative limits. Accordingly, areas with high cumulative pollutants' concentrations or nitrate concentration exceeding regulatory limits should be properly identified by the labels of either moderate or high vulnerability. On the other hand, areas where pollutant concentrations do not exceed regulatory limits or low nitrate concentrations should be related with either a low or null presence of pollutant sources and/or a low vulnerability of the aquifer.

Comparison of the cumulative map in Fig. 3 with the DRASTIC vulnerability map in Fig. 4 shows that this vulnerability index is capable of identifying some areas of high cumulative exposure but fails in others probably due to the inability of the linear nature of equation (1) to properly resolve class and/or scale changes. Liggett and Allen (2011) evaluated the sensitivity of DRASTIC by using different data sources, interpretations and mapping approaches. They demonstrated that small-scale changes in vulnerability are not properly identified by the original DRASTIC methodology. The results in Fig. 4 are complemented with the frequencies listed in Table 3 for any of the screened pollutants (Al, Sb, Ar, Cd, Cr, Mo, Ni, Fe, Mn, Se, Ba, Cu, Pb, Zn, Mg, nitrate, nitrite and sulfates) exceeding cumulative legal threshold by year and vulnerability class. This table includes statistics for only the NO_3 data. Results reveal that the areas classified as "low vulnerability" include several measurement stations exceeding legal limits. In particular, Table 3 indicates that 10 points located in the "low vulnerability" class exceeded the limit for NO_3 during 2002. Also, it should be noticed that the average NO_3 concentrations are not consistent in the low, moderate and high vulnerability areas.

The above inconsistent results for vulnerability categorization using DRASTIC indicate that there is a need for an improved groundwater vulnerability approach that integrates the management of all georeferenced data and the calculation of a reliable vulnerability index for risk assessment. Self-organizing maps can provide the foundations for such an integrated approach. Two SOM approaches have been developed. The first consists of a DRASTIC-based SOM model that uses the original seven DRASTIC parameters for training the map with parameters rated as in the DRASTIC procedure. The trained map is used thereafter to estimate the intrinsic vulnerability. The second, is a specific vulnerability SOM model that self-organizes the six raw hydrogeological and climate data [piezometric level (H), annual rainfall (P), soil's permeability (Ks), land

surface slopes (%s), aquifer media (A), hydraulic conductivity of the aquifer (Ka)] that generate the seven DRASTIC features in Equation (1), together with a layer of land use information (land). The land use has been included in the proposed specific vulnerability SOM model because it has been demonstrated to influence the vulnerability analysis (Secunda et al., 1998). It should be noted that the addition of land use information implicitly incorporates specific stressor/potential pollution information into the vulnerability analysis (Chen et al., 2010). This addition converts any intrinsic vulnerability index, such as DRASTIC, into a specific vulnerability index to assess groundwater quality (Martinez-Bastida et al., 2010) by providing information regarding specific sources of potential contamination.

3.2 Intrinsic SOM-based vulnerability with original DRASTIC variables

To evaluate first the applicability of SOM for groundwater vulnerability assessment, a DRASTIC-based SOM model was developed by self-organizing the seven DRASTIC features in Equation (1) to generate intrinsic vulnerability maps. Thus, the input of this model was a vector with the DRASTIC features as elements. The current methodology improves previous SOM-based approach (Pistocchi et al., 2011) by using (i) a new formulation for the vulnerability index, (ii) a new weighting schema, and (iii) a double SOM clustering. The weighting scheme uses the DRASTIC weighting (Aller et al., 1987) as a mask in the input of the SOM to perform the training of the map. The vulnerability index (*vIndex*) is defined by:

$$vIndex_j = \sqrt{\frac{\sum v_{ij}^{\,2}}{n}} \tag{5}$$

where v_{ij} is the value of variable i in the prototype vector of cluster j. Similarly, a weighted index, *wIndex*, based on normalized DRASTIC weights (i.e., weights are normalized so that their summation is equal to one) can be defined as:

$$wIndex_j = \sum w_i v_{ij} \tag{6}$$

A double SOM scheme was also developed to refine the estimation of the vulnerability index. The double SOM algorithm consists of the sequential application of two SOM layers, the output of the first SOM is used as the input of the second SOM. Finally, the index obtained using either Eq. (5) or Eq. (6) was assigned to low, moderate or high vulnerability categories by assuming a normal distribution of its values and using the 48% and 68% percentiles as the cut-off criteria for categorization (Aller et al., 1987).

An intrinsic vulnerability map at the local scale of the Camp de Tarragona was developed using the above methodology (Figure 5). Briefly, the DRASTIC-based SOM was developed using the double-SOM approach with normalized DRASTIC weights in the first SOM clustering and the *vIndex* calculation using the centroids of the second SOM. Different configurations were evaluated to optimize the number of units in each SOM layer. The optimal architecture of the double SOM consisted of 2542 and 260 units for the

first and second map, respectively. Following the methodology used in the previous section, the intrinsic vulnerability map was validated using NO_3 concentrations and cumulative maps of pollutants as groundwater quality indicators.

Statistics of NO_3 and cumulative exposure of pollutants for the DRASTIC-based SOM vulnerability model presented in Table4 show a significant improvement in the vulnerability classification relative to the previous DRASTIC results (Table 3). For the complete set of pollutants, only 3 measurement stations located in areas of *low* vulnerability exceeded regulatory thresholds (Table 4). Figure 5 (left) depicts the clustering structure (u-matrix) and the distribution of variables over the SOM space (c-planes) for the seven DRASTIC parameters. The structure of the u-matrix indicates the presence of well-defined clusters (dark blue regions) that confirm that the SOM approach is able to identify regions with similar vulnerability patterns. The inspection of the c-planes (Figure 5, left) shows that the D and I features are correlated (i.e., have a similar distribution over the SOM). Thus, features D and I can be considered as redundant and only one of the two should be included in the vulnerability analysis.

The area of *low* vulnerability identified by the DRASTIC-based SOM model (Figure 5, right) is very small. The analysis of the NO_3 statistics indicates, in this area, low concentration values, with a mean 24.28 mg/L, without any concentration value exceeding the legal limit for NO_3(Table 4). The correct identification of “moderate” and “high” vulnerability areas according to nitrate measures and cumulative pollutant data confirms that the non-linear clustering capabilities of the SOM captures better the relationship between the seven DRASTIC features and vulnerability than it does the simple aggregation method of the original DRASTIC.

3.3SOM-based specific vulnerability model

The SOM capabilities to generate vulnerability maps that avoid some of the *ad hoc* rating steps involved in the DRASTIC methodology were also studied. To evaluate this approach, a SOM-based vulnerability model was developed using the primary (raw) hydrogeological data that generate the seven DRASTIC features combined with information regarding the land uses (land). The raw values of numerical variables such as piezometric level, annual rainfall, and land surface slopes were used without applying the expert criteria rating process used in the DRASTIC methodology. Categorical variables, like hydraulic conductivity, soil permeability and land uses, were rated as in DRASTIC. Before developing the SOM, all input variables were normalized in the range [0,1]with respect to predefined upper and lower limits (Table 5). The upper limit is the highest value that a given parameter can reach at a regional scale (i.e., regional maximum value). Values greater than the limit are considered to have the same impact on vulnerability. Likewise, the regional minimum corresponds to the lowest value that a parameter can reach and values lower than the limit are considered with the same impact on vulnerability. The data normalization process ensures the global applicability of the current vulnerability index.

The SOM-based approach used the double SOM architecture described in the previous section. The optimal sizes for each SOM structure were 2535 and 260 units for the first and

second map, respectively. Figure 6(left) shows the u-matrix and the distribution of the seven selected variables over the SOM lattice (c-planes). The corresponding SOM vulnerability map is depicted in Fig. 6 (right). The inspection of the c-planes in Fig. 6(left) indicates that there is no correlation between input variables (i.e., different distributions over the SOM space). Table 6 presents the statistics of the point cumulative map and NO_3 concentration for the SOM-based vulnerability map. The "low" vulnerability zone is also small like in the previous DRASTIC-based SOM model. In the SOM-based approach, however, monitoring stations located in areas of low vulnerability do not exceed the legal threshold for any pollutants. A total of 40% and 57% of the monitoring stations exceeded the regulatory limits for NO_3 in areas with "moderate" and "high" vulnerabilities, respectively.

The SOM-based approach yields a specific vulnerability map that (i) provides a more detailed distinction between vulnerability zones within the area considered since they stem from the labeling of directly identified hydrogeological-climate classes, and (ii) correlates better with the cumulative map for combined effects of pollutants shown in Fig. 3 than the previous DRASTIC and DRASTIC-based SOM maps presented in Fig. 4 and Fig. 5 (right), respectively. The improvements arise from the use of a non-linear classification algorithm such as SOM to identify hydrogeological classes and to better relate them with vulnerability. Also, the adoption of a more appropriate set of parameters to categorize the domain and to assess vulnerability in the area considered is a key factor that improves vulnerability predictions. It should be noted that SOM preserves the topology of data and, even though geographical coordinates were omitted in the training process, the continuity of vulnerable areas (classes) in the physical domain is also maintained in the projection.

The SOM-based approach for groundwater quality (specific vulnerability) assessment using self-organizing maps yields reliable specific vulnerability maps from properly hydrogeological and climate variables, thus providing a sound basis for consistent land use planning and water resources management policies.

4. Regional scale vulnerability assessment

The SOM-based vulnerability methodology described in the previous section has been applied at the regional scale of Catalonia. While the number of measurement stations is usually enough at the local scale to yield accurate spatial interpolations, this is not always the case when considering wider regions, where some areas cannot be covered by the measurement network. To analyze data at the regional level an up-scaling process capable of spanning local data onto a wider region while maintaining spatial relationships is required. The intrinsic topology preservation properties of the SOM are exploited here to perform this interpolation and up-scaling process and to complete the development of the current integrated approach for vulnerability assessment.

4.1 Cumulative pollution maps

Cumulative maps for different contaminants were developed using two approaches. First, geostatistics were used to create continuous surfaces from spatial sample measurements. Ordinary and simple kriging were used to generate concentration maps of

the pollutants. Second, the SOM was used to generate smooth maps by mimicking the cokriging technique with the hydrogeological units (Figure 1) as constraints for the interpolation process.

Figure 7 compares two nitrate concentration maps for year 2002 estimated using kriging interpolation and SOM interpolation with the hydrogeological unit constraint. The kriging interpolation (Figure 7, top) is governed mainly by the shape of the distribution model used to build the variogram. In contrast, SOM-based interpolation (Figure 7, bottom) yields a less regular response due to the effect of the inclusion of the spatial information concerning the hydrogeological units. In this later case, the interpolation process is mainly governed by the variables which constrain the modeled exposure. Even though the SOM-based interpolation results in higher estimates relative to kriging (Figure7), the spatial location of the main hot spots is preserved. Smoother exposure nitrate concentration distributions can be obtained by averaging the concentration values given by the SOM units located in the direct neighborhood of the best matching unit. The main advantage of proposed SOM interpolation as a cokriging technique for pollutants' concentrations in aquifers is that it is geologically consistent due to the fact that each hydrogeological area constitutes an independent geological unit, i.e., there is no water flow continuity from one unit to the other. On the other hand, the consistent application of the kriging technique to each hydrogeological area requires enough geo-referenced data at each unit (statistically at least 10-15 points), which is impossible with the available data. The SOM based method requires less data for producing smooth estimations of exposure concentrations.

Exposure maps have been generated with both interpolation methodologies for the whole data set of 17 pollutants considered. Their cumulative effects have been studied with cumulative exposure maps to identify elevated risk areas where high concentrations of contaminants converge. To build these maps the same procedure used for the local vulnerability analysis was followed, i.e., concentration maps were first converted to Boolean data (1's and 0's) indicating if a grid cell exceeded regulatory threshold values or not, respectively, and afterwards, a Boolean intersection method for aggregation was used to generate the cumulative maps. Figure 8 depicts the resulting cumulative maps for year 2002 obtained using kriging and SOM interpolation techniques. At year 2002, fourteen pollutants (NO_2, NO_3, SO_4, Fe, Mn, Al, Sb, Ar, Ba, Mo, Ni, Pb, Se, Zn) exceeded the legal threshold in at least one location in Catalonia. Results range from 1 (Low) to 6 (High) in areas were one or more pollutants exceeded their legal threshold. It can be observed that in almost all Catalonia (Figure 8) at least one pollutant exceeded the legal threshold and a few hot spots of 6 pollutants exceeding legal threshold are also present.

Inspection of these cumulative maps shows again that the response of the kriging-based maps is mainly driven by the distribution function model assumed for the variogram. Again the spatial location of hotspots is consistent in these two map representations. The cumulative map obtained by the combination of single pollutant interpolated maps using SOM gives slightly higher estimates than the one obtained using kriging.

4.2 DRASTIC intrinsic vulnerability model

DRASTIC vulnerability maps have been generated and used as a reference to evaluate the current SOM approach at the regional scale of Catalonia. The same methodology previously used at the local scale was applied at a regional level to calculate DRASTIC features (Piscopo, 2001). Depth to water table (D) was generated from piezometric data. Net recharge layer (R) was obtained from the addition of surface slopes, annual rainfall and soil´s permeability layers. The aquifer media (A) and hydraulic conductivity (C) features were generated from aquifer permeability data. Soil media layer (S) was obtained from infiltration capacity data. Topography (T) was generated using the DTM. Impact to vadose zone layer (I) was generated from the addition of soil`s permeability and the depth to water table layer. The DRASTIC vulnerability map for Catalonia (Figure 9) included the three vulnerability categories of low, moderate and high. The most vulnerable areas (Figure 9) are located in the east part of Catalonia near the Mediterranean Sea.

Table 7 presents pollutants statistics for the regional scale vulnerability areas identified by DRASTIC. The analysis of NO_3 pollution data reveals that “low” vulnerability class has higher mean concentration value than “moderate” and “high” zones (47.73 mg/l for the “low” class but 40.02 and 35.22 mg/l for “moderate” and “high” classes). Also, the “low” vulnerable class includes the 47% of monitoring stations exceeding regulatory limits for all pollutants considered in the study. These results confirm, at the regional scale of Catalonia, the drawbacks of the DRASTIC methodology discussed in the previous section.

4.3 SOM-based specific vulnerability model

The SOM-based vulnerability methodology was applied to the regional scale of Catalonia. A vulnerability map at the regional scale was generated by self-organizing the seven hydrogeological and climate properties [piezometric level (H), annual rainfall (P), soil`s permeability (Ks), aquifer media (A), land surface slopes (%s), hydraulic conductivity of the aquifer (Ka) and land uses (land)]. The optimal clustering was obtained using a double-SOM of 4480 and 345 units, respectively. Figure 10 shows the three classes obtained using these variables in the training process of the SOM for Catalonia.

The vulnerability map obtained via the SOM approach is consistent with cumulative exposure maps (Figure 8). In addition, the SOM based vulnerability index produces vulnerability maps which are compatible with those produced using the DRASTIC index. Nevertheless, the SOM-based vulnerability method is capable of coping with the scarcity of environmental data sets. The SOM provides a consistent framework to infer the missing data and to estimate (interpolate) the exposure concentration of diverse pollutants even from few data. It should be noted, however, that if the amount of information is insufficient the quantization error (Eq. 3) provides a measure to diagnose the reliability of the SOM interpolation.

Visual comparison of Figure 9 (DRASTIC vulnerability) and Figure 10 (current *SOM* approach) indicates that the SOM vulnerability index yields lower vulnerability estimates but with a more consistent spatial continuity than for the DRASTIC model. This is due to the clustering and topology preservation properties of the SOM learning algorithm. The aquifers located at the North-East part of Catalonia are the most vulnerable mainly due to

the impact of the high population density areas that are primarily located along the Mediterranean coast. Other vulnerable areas, related to human activities, are located near the most populated cities in Catalonia where most of the Catalan industry and logistic distribution centers are also located.

Statistics of pollutant concentrations for the vulnerability classes detected in the SOM-based vulnerability map for Catalonia are presented in Table 8. SOM-based vulnerability classes are highly correlated with NO_3 mean concentrations. The mean values increase with vulnerability and the "high" vulnerability zone has an average NO_3 concentration that surpasses the regulatory limit. The analysis of cumulative point data reveals that the "low" vulnerability class includes only 3% of measurement stations exceeding regulatory limits for the area of study.

Comparison of SOM-based vulnerability maps for Camp de Tarragona obtained at the local scale (Figure 11, right) and at the regional scale (Figure 11, left) shows good agreement after the up-scaling process. The regional scale model provides a first estimate of vulnerable areas that can be further evaluated using local scale models with a more refined data resolution.

5. Conclusions

The Self-Organizing Map algorithm provides a consistent framework that can be successfully applied at different levels of the ERA framework and at different spatial resolutions. The proposed methodology for groundwater vulnerability assessment has been tested at two different scales. At the local level in the Camp de Tarragona hydrogeological unit, where the concentration distribution functions of diverse stressors have been determined from available data, the clustering capabilities of the SOM provide vulnerability estimates without requiring previous expert's knowledge ratings of numerical variables. At wider scales such as the regional level of Catalonia, the SOM has been successfully used to deal with missing data, spatial interpolation, probabilistic risk analysis, and intrinsic and specific vulnerability assessment. This latter aspect is addressed by the development of a new vulnerability index which is independent of expert's criteria and is able to integrate several sources of diverse hydrogeological and climatic information. This vulnerability index is standardized and discretized in such a way that it is independent of the scale of the geographical region considered. The values obtained from this new index can thus be easily applied to estimate groundwater vulnerability at different scales within Europe. The new methodology provides more detailed and geographically consistent vulnerability maps than the well established DRASTIC methodology in all its variations proposed previously in the literature.

The SOM approach can help regulators and policy makers to understand the relationships between the potential stressors of concern in an environmental risk scenario such as the pollution of groundwater and the vulnerability of drinking water sources. As the proper management of water resources is becoming a major concern in Europe and in the rest of the world, the proposed SOM based approach for groundwater intrinsic and

specific vulnerability assessment provides a reliable and adaptable tool for resource planning and decision making.

Future extensions to this technique should provide mechanisms to automatic tuning the optimal neighborhood used for exposure modeling. Also, the categorization of the vulnerability classes could be exploited to produce vulnerability-driven risk assessment models in which accurate probabilistic risk models could be adjusted to each vulnerability category. Additionally, the use of hierarchical ensembles of SOMs could provide an integrated view of vulnerability at different spatial scales and facilitate the inference of relationships between vulnerability estimates at each scale.

Conflict of interest

The authors do not have competing financial interests to declare.

Acknowledgments

This research was financially partly supported by the European Union (NoMiracle Project, European Commission, FP6 Contract No. 003956), the Spanish Ministry of Science and Innovation (MICINN, CTM2011-24303), the Generalitat de Catalunya (2009SGR-1529) and partly by the Servei de Gestió de la Recerca and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of Generalitat de Catalunya. Francesc Giralt acknowledges the support received from the *Distinció a la Recerca* (Generalitat de Catalunya).

References

Ahmed, A. A., 2009. Using Generic and Pesticide DRASTIC GIS-based models for vulnerability assessment of the Quaternary aquifer at Sohag, Egypt. Hydrogeol. J. 17, 1203-1217.

Aller, L., Lehr, J.H., Petty, R., Bennett, T., 1987.Drastic: A standardized system to evaluate groundwater pollution potential using hydrogeologic settings. US EPA. Proj. Summ. EPA/600/S2-87/035.

Almasri, M. N., 2008.Assessment of intrinsic vulnerability to contamination for Gaza coastal aquifer, Palestine. J. Environ. Manage. 88, 577-593.

Andreo, B., Goldscheider, N., Vadillo, I., Vías, J.M., Neukum, C., Sinreich, M., Jiménez, P., Brechenmacher, J., Carrasco, F., Hötzl, H., Perles, M.J., Zwahlen, F., 2005. Karst groundwater protection: First application of a Pan-European Approach to vulnerability, hazard and risk mapping in the Sierra del Líbar (Southern Spain). Sci. Total Environ. 357, 54-73.

Babiker, I.S., Mohamed, M.A.A., Hiyama, T., Kato, K., 2005. A GIS-based DRASTIC model for assessing aquifer vulnerability in Kakamigahara Heights, Gifu Prefecture, central Japan. Sci. Total Environ. 345, 127-140.

Bojórquez-Tapia, L.A., Cruz-Bello, G.M., Luna-González, L., Juárez, L., Ortiz-Pérez, M. A., 2009. V-DRASTIC: Using visualization to engage policymakers in groundwater vulnerability assessment. J. Hydrol. 373(1-2), 242-255.

Burchart, A., Leppig, B., MacDonald, A., Müller, B., Wimmer, G., 2006. Mapping the groundwater vulnerability in North Rhine-Westphalia, Germany. Environ. Eng. Sci. 23, 574-578.

Céréghino, R., Park, Y. S., 2009. Review of the Self-Organizing Map (SOM) approach in water resources: Commentary. Environ. Modell. Softw. 24, 945-947.

Chen, Y., Yang, S., Dong, S., Li, Y., Sun, B., Shao, Z., 2010. Influence of Agricultural Activity and Aquifer Intrinsic Vulnerability on Groundwater Quality in the Dagu River Watershed (Qingdao, China).4th Int. Conf.Bioinform. Biomed. Eng. (iCBBE).1-6, 18-20.

Civita, M., 1994. Le carte della vulnerabilità degli acquiferi all inquinamento: Teoria and practica. Pitagora Editrice, Bologna.

Dixon, B., 2005a. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: A GIS-based sensitivity analysis. J. Hydrol. 309, 17-38.

Dixon, B., 2005b. Groundwater vulnerability mapping: A GIS and fuzzy rule based integrated tool. Appl. Geogr. 25, 327-347.

Draoui, M., Vias, J., Andreo, B., Targuisti, K., 2008. A comparative study of four vulnerability mapping methods in a detritic aquifer under Mediterranean conditions. Environ. Geol. 54, 455-463.

Duijvenbooden, W.V., Waegeningh, H.G.V., 1987.Vulnerability of Soil and Groundwater to Pollutants. P. Int. Conf. Vulnerability of Soil and Groundwater to Pollutants. Delft, The Netherlands.

Gemitzi, A., Petalas, C., Tsihrintzis, V.A., Pisinaras, V., 2006. Assessment of groundwater vulnerability to pollution: a combination of GIS, fuzzy logic and decision making techniques. Environ. Geol. 49, 653-673.

Goldscheider, N., 2005. Karst groundwater vulnerability mapping: application of a new method in the Swabian Alb, Germany. J. Hydrol. 13, 1431-2174.

Hu, K., Huang, Y., Li, H., Li, B., Chen, D., White, R.E., 2005. Spatial variability of shallow groundwater level, electrical conductivity and nitrate concentration, and risk assessment of nitrate contamination in North China Plain. Environ. Int. 31, 896-903.

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modeling and application. Environ. Modell. Softw. 23, 835-845.

Kangas, J.A., Kohonen, T.K., Laaksonen, J.T., 1990. Variants of self-organizing maps. IEEE T. Neural Network. 1(1), 93-99.

Kaski, S., Venna, J., Kohonen, T., 1999. Coloring that reveals high dimensional structures in data. In T. Gedeon, P. Wong, S. Halgamuge, N. Kasabov, D. Nauck, K. Fukushima, editors, P. on ICONIP'99, 6th Int. Conf. Neural Inform. Process. II, 729-734.

Kazemi, S.M., Hosseini, S.M., 2011. Comparison of spatial interpolation methods for estimating heavy metals in sediments of Caspian Sea. Expert Syst. Appl. 38(3), 1632-1649.

Kaski, S., 1997. Data exploration using self-organizing maps. Department of Computer Science and Technology. Doctor of Technology. Helsinki University of Technology.

Kohonen, T., 1990. The self-organizing map. Neurocomputing. 21, 1-6.

Kohonen, T., 2001. Self-Organizing Maps. Springer-Verlag, Berlin.

Lahr, J., Kooistra, L., 2009. Environmental risk mapping of pollutants: State of the art and communication aspects. Sci. Total Environ. 408(18), 3899-3907.

Laine, S., 2003. Using visualization, variable selection and feature extraction to learn from industrial data. Department of Computer Science and Technology. Dissertation for the degree of Doctor of Technology. Helsinki University of Technology.

Liggett, J. E. Allen, D., 2011. Evaluating the sensitivity of DRASTIC using different data sources, interpretations and mapping approaches. Environ. Earth Sci. 62, 1577-1595.

Lindström, R., 2005. Groundwater vulnerability assessment using process-based models. Dissertation for the Degree of Doctor in Technology. Helsinki University of Technology.

Mackay, D., Di Guardo, A., Paterson, S., Kicsi, G., Cowan, C.E., Kane, D.M., 1996. Assessment of chemical fate in the environment using evaluative, regional and local-scale models: Illustrative application to chlorobenzene and linear alkylbenzenesulfonates. Environ. Toxicol. Chem. 15(9), 1638-1648.

Mao, Y.-Y., Zhang, X-G., Wang, L-S., 2006. Fuzzy pattern recognition method for assessing groundwater vulnerability to pollution in the Zhangji area. J. Zhejiang Univ.Sci. A. 7, 1917-1922.

Martínez-Bastida, J.J., Arauzo, M., Valladolid, M., 2010. Intrinsic and specific vulnerability of groundwater in central Spain: the risk of nitrate pollution. Hydrogeol. J. 18(3), 681-698.

Martínez-Santos, P., Llamas, M.R., Martínez-Alfaro, P.E., 2008. Vulnerability assessment of groundwater resources: A modeling-based approach to the Mancha Oriental aquifer, Spain. Environ. Modell. Softw. 23, 1145-1162.

Masetti, M., Sterlacchini, S., Ballabio, C., Sorichetta, A., Poli, S., 2009. Influence of threshold value in the use of statistical methods for groundwater vulnerability assessment. Sci. Total Environ. 407, 3836-3846.

Mazari Hiriart, M., Cruz Bello, G., Bojórquez Tapia, L.A., Juárez Marushi, L., Alcantar López, G., Marín, L.E., Soto Galera, E., 2003. Groundwater vulnerability assessment for organic compounds: Fuzzy multicriteria approach for Mexico City. Environ. Manage. 37, 410-421.

Mishima, Y., Takada, M., 2011. Evaluation of intrinsic vulnerability to nitrate contamination of groundwater: appropriate fertilizer application management. Environ. Earth Sci. 63, 571-580.

Moore J. S., 1990. SEEPAGE: A system for early evaluation of the pollution potential of agricultural groundwater environments. USDA. SCS, Northeast Technical Center. Geol. Technol. Note.

Neukum, C., Hötzl, H., Himmelsbach, T., 2008. Validation of vulnerability mapping methods by field investigations and numerical modeling. Hydrogeol. J. 16, 641-658.

Nolan, B.T., Hitt, K., 2006. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. Environ. Sci. Technol. 40, 7834-7840.

Nolan, B.T., Ruddy, B.C., Hitt, K., Helsel, D.R., 1997. Risk of nitrate in groundwaters of the United States - A national perspective. Environ. Sci. Technol. 31, 2229-2236.

Passuello, A., Cadiach, O., Perez, Y., Schuhmacher, M., 2012. A spatial multicriteria decision making tool to define the best agricultural areas for sewage amendment. Environ. Int. 38(1), 1:9.

Panagopoulos, G.P., Antonakos, A. K., Lambrakis, N. J., 2006. Optimization of the DRASTIC method for groundwater vulnerability assessment via the use of simple statistical methods and GIS. Hydrogeol. J. 14, 894-911.

Peeters L., Bacao, F., Lobo, V., Dassargues, A., 2007. Exploratory data analysis and clustering of multivariate three-dimensional spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map. Hydrol. Earth Syst. Sci. 11, 1309-1321.

Perles Roselló, M.J., Vías Martínez, J.M., Andreo Navarro, B., 2009. Vulnerability of human environment to risk: case of groundwater contamination risk. Environ. Int. 35(2), 325-335.

Piscopo, G., 2001. Groundwater vulnerability map explanatory notes. Center of Natural Resources. NWS Department of Land and Water Conservation. Parramatta.

Pistocchi, A., Groenwold, J., Lahr, J., Loos, M., Mujica, M., Ragas, A.M.J., Rallo, R., Sala, S., Schlink, U., Strebel, K., Vighi, M., Vizcaino, P., 2011. Mapping cumulative environmental risks: examples from the EU NoMiracle project. Environ. Model. Assess. 16, 119-133.

Pons, X., 2006. MiraMon Geographic Information System and Remote sensing software. http://www.creaf.uab.es/miramon/

Popescu, I.C., Gardin, N., Brouyere, S., Dassargues, A., 2008. Groundwater vulnerability assessment using physically based modeling: from challenges to pragmatic solutions. P. of Calibration and Reliability in Groundwater Modeling: Credibility in Modeling, ModelCARE 2007 Copenhague, Denmark 320, 83-88.

Rallo, R., 2007. Multi-tier framework for the inferential measurement and data-driven modeling. PhD dissertation. Universitat Rovira i Virgili. Spain.

Rahman, A., 2008. A GIS based DRASTIC model for assessing groundwater vulnerability in shallow aquifer in Aligarh, India. Appl. Geogr. 28, 32-53.

Sanchez-Martos, F., Aguilera, P.A., Garrido-Frenich, A., Torres, J.A., Pulido-Bosch, A., 2002. Assessment of groundwater quality by means of self-organizing maps: application in a semiarid area. Environ. Manage. 30, 716-726.

Secunda, S., Collin, M.L., Melloul, A.J., 1998. Groundwater vulnerability assessment using a composite model combining DRASTIC with extensive agricultural land use in Israel's Sharon region. J. Environ. Manage. 54(1), 39-57.

Sinan, M., Razack, M., 2009. An extension to the DRASTIC model to assess groundwater vulnerability to pollution: application to the Haouz aquifer of Marrakech (Morocco). Environ. Geol. 57, 349-363.

Stigter, T.Y., Ribeiro, L., Carvalho Dill, A.M.M., 2006. Evaluation of an intrinsic and specific vulnerability assessment method in comparison with groundwater Stalinization and nitrate contamination levels in two agricultural regions in the south of Portugal. Hydrogeol. J. 14, 79-99.

Tilahun, K., Merkel, B.J., 2010. Assessment of groundwater vulnerability to pollution in Dire Dawa, Ethiopia using DRASTIC. Environ. Earth Sci. 59, 1485-1496.

Tutmez, B., Hatipoglu, Z., 2010. Comparing two data driven interpolation methods for modeling nitrate distribution in aquifer. Ecol. Inform. 5(4), 311-315.

Umar, R., Ahmed, I., Alam, F., 2009. Mapping groundwater vulnerable zones using modified DRASTIC approach of an alluvial aquifer in parts of Central Ganga plain, western Uttar Pradesh. Journal of the Geological Society of India 73(2): 193-201.

Uricchio, V.F., Giordano, R., Lopez, N., 2004. A fuzzy knowledge-based decision support system for groundwater pollution risk evaluation. J. Environ. Manage. 73, 189-197.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE T. Neural Networ. 11(3), 586-600.

Vías J., Andreo, B., Raubar, N., Hötzl, H., 2010. Mapping the vulnerability of groundwater to the contamination of four carbonate aquifers in Europe. J. Environ. Manage. 91(7), 1500-1510.

Villa, F., McLeod, H., 2002. Environmental Vulnerability Indicators for Environmental Planning and Decision-Making: Guidelines and Applications. Environ. Manage. 29, 335-348.

Worrall, F., Besien, T., 2005. The vulnerability of groundwater to pesticide contamination estimated directly from observations of presence or absence in wells. J. Hydrol. 303, 95-107.

Worrall, F., Besien, T., Kolpin, D. W., 2002. Groundwater vulnerability: interactions of chemical and site properties. Sci. Total Environ. 299(1-3), 131-143.

Worrall, F., Kolpin, D. W., 2003. Direct assessment of groundwater vulnerability from single observations of multiple contaminants. Water Resour. Res. 39, 1345-1345.

Zabeo, A., Pizzol, L., Agostini, P., Critto, A., Giove, S., Marcomini, A., 2011. Regional risk assessment for contaminates sites Part 1: Vulnerability assessment by multicriteria decision analysis. Environ. Int. 37, 1295-1306.

Zare, A.H., Bayat, V. M., Daneshkare, A.P., 2011. Forecasting nitrate concentration in groundwater using artificial neural network and linear regression models. Int. Agrophys. 25, 187-192.

Table 1

Sources and resolution of hydrogeological data

Data	Layer type (Original resolution)	Source
Geological	Raster (200 by 200 meters)	Departament de Medi Ambient i Habitatge, Catalunya
Land use	Raster (30 by 30 meters)	Departament de Medi Ambient i Habitatge, Catalunya
Annual rainfall	Raster (200 by 200 meters)	Departament de Medi Ambient i Habitatge, Catalunya
Digital terrain model	Raster (200 by 200 meters)	Institut Cartogràfic de Catalunya
Aquifer's permeability	Raster (1000x1000 meters)	Instituto Geológico y Minero de España
Piezometrics level	Point data (559 measures)	Agencia Catalana de l'Aigua (ACA)
Soil's permeability	Point data (123 measures)	Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas de España (CIEMAT).

Table 2

Statistics of heavy metals and pesticides in groundwater at year 2002

	Regional scale: Catalonia					Local scale: Camp de Tarragona				
	N	min	max	mean	Nexd	N	min	max	mean	Nexd
NO_2 (mg/l)	637	0.30	296.30	42.48	208	100	0.70	168.50	46.88	42
NO_3 (mg/l)	637	0.02	2.04	0.07	9	100	0.02	1.33	0.072	3
SO_4 (mg/l)	390	5.00	2704.00	198.00	91	68	30.50	496.00	138.62	11
Fe (µg/l)	401	20.00	82830.00	703.72	82	56	20.00	10680.00	599.2	14
Mn (µg/l)	401	5.00	5006.00	103.02	73	56	5.00	1048.00	79.57	9
Al (µg/l)	401	60.00	516.00	66.97	5	56	60.00	194.00	64.79	0
Sb (µg/l)	401	4.00	12.00	4.06	4	56	4.00	10.00	4.12	1
As (µg/l)	401	4.00	163.00	4.74	10	56	4.00	29.00	4.70	2
Ba (µg/l)	401	0.20	1020.00	81.61	6	56	11.7	1020.00	115.72	3
Cd (µg/l)	401	0.50	4.70	0.54	0	56	0.50	1.10	0.51	0
Cu (µg/l)	401	3.00	460.00	10.31	0	56	3.00	148.00	7.75	0
Cr (µg/l)	401	4.00	14.00	4.10	0	56	4.00	6.00	4.07	0
Mo (µg/l)	420	1.00	31.00	1.72	401	56	1.00	5.00	1.30	56
Ni (µg/l)	420	5.00	109.00	6.40	10	56	5.00	55.00	6.50	1
Pb (µg/l)	401	5.00	99.00	6.01	16	56	5.00	99.00	6.88	1
Se (µg/l)	401	3.00	108.00	10.20	125	56	3.00	22.00	6.91	7
Zn (µg/l)	401	4.00	7051.00	132.20	3	56	4.00	2600.00	129.84	0

N: number of measures; min: minimum value; max: maximum value; mean: arithmetic mean; Nexd: number of exceeded threshold values

Table 3.

Frequency of exceeding cumulative legal threshold and nitrate concentrations statistics by vulnerability class in DRASTIC-Aller vulnerability map for Camp de Tarragona area

Cumulative				Nitrate			
#Exd	Low	Mod	High	Stats	Low	Mod	High
0	12	29	1	N	31	58	10
1	15	18	3	Nexd	10	25	6
2	5	18	3	mean	40.46	49.04	49.55
3	2	1	1	min	2.3	0.7	0.9
4	2	1	2	max	167.5	168.5	113
5	1	1	1	std	34.85	37.07	33.14
6	0	2	0	median	33.35	45.6	59.75
				Q1	15.35	20.3	23.5
				Q3	63.6	66.95	64.6

Exd: number of pollutants exceeding legal thresholds in the same monitoring station. N: number of NO3 measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

Table 4.

Frequency of exceeding cumulative legal threshold and nitrate concentrations statistics by vulnerability class in DRASTIC-based SOM vulnerability map for Camp de Tarragona area

Cumulative				Nitrate			
#Exd	Low	Mod	High	Stats	Low	Mod	High
0	0	30	12	N	3	66	30
1	2	26	8	Nexd	0	21	20
2	1	13	12	mean	24.28	40.67	61.24
3	0	2	2	min	18.8	0.7	0.9
4	0	2	3	max	33.25	167.5	168.5
5	0	0	2	std	7.91	33.3	38.88
6	0	3	0	median	20.7	33.8	63.25
				Q1	18.8	16.85	36.4
				Q3	33.35	60.1	78.2

Exd: number of pollutants exceeding legal thresholds in the same monitoring station. N: number of NO_3 measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

Table 5.

Regional maximum and minimum for groundwater vulnerabilities features for the SOM-based vulnerability model

Variable	Maximum	Minimum
Piezometric level (H)	50 m	0 m
Annual rainfall (P)	2000 mm	70 mm
Soil's permeability (Ks) *	10	1
Land surface slopes (%s)	60%	0%
Aquifer media (A) *	10	1
Hydraulic conductivity of aquifer (Ka) *	10	1
Soil Uses (land) *	10	1

* Categorical variables were considered using DRASTIC ratings

Table 6.

Frequency of exceeding cumulative legal threshold and nitrate concentrations statistics by vulnerability class in SOM-based vulnerability map for Camp de Tarragona area

Cumulative				Nitrate			
#Exd	Low	Mod	High	Stats	Low	Mod	High
0	0	37	5	N	0	85	14
1	0	31	5	Nexd	-	33	8
2	0	22	4	mean	-	46.25	47.33
3	0	4	0	min	-	0.7	0.9
4	0	3	2	max	-	168.5	73.2
5	0	1	1	std	-	37.66	23.42
6	0	3	0	median	-	36.4	55.86
				Q1	-	18.8	33.5
				Q3	-	66.95	64.9

Exd: number of pollutants exceeding legal thresholds in the same monitoring station. N: number of NO_3measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

Table 7.

Frequency of exceeding cumulative legal threshold and nitrate concentrations statistics by vulnerability class in DRASTIC vulnerability map for Catalonia area

Cumulative				Nitrate			
#Exd	Low	Mod	High	Stats	Low	Mod	High
0	126	55	57	N	317	172	148
1	128	62	32	Nexd	117	50	41
2	50	53	26	mean	47.73	40.02	35.22
3	35	26	33	min	0.3	0.3	0.3
4	16	11	17	max	293.7	202.67	296.3
5	5	2	8	std	48.03	40.66	41.84
6	3	2	0	median	33.9	25.35	19.71
				Q1	12.3	9.28	5.15
				Q3	68.2	58.02	53.22

Exd: number of pollutants exceeding legal thresholds in the same monitoring station. N: number of NO_3measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

Table 8.

Frequency of exceeding cumulative legal threshold and nitrate concentrations statistics by vulnerability class in SOM-based specific vulnerability map for Catalonia area

Cumulative				Nitrate			
#Exd	Low	Mod	High	Stats	Low	Mod	High
0	5	173	60	N	16	462	159
1	12	167	43	Nexd	3	136	69
2	1	95	33	mean	33.29	39.31	53.67
3	1	62	31	min	6.1	0.3	0.3
4	0	28	16	max	107.1	293.7	296.3
5	1	7	7	std	27.62	41.81	52.98
6	0	4	1	median	24.45	25.03	43.07

	Q1	14.1	8.5	12.3
	Q3	43.5	58.2	77.05

Exd: number of pollutants exceeding legal thresholds in the same monitoring station. N: number of NO_3measures; Nexd: number of exceeded threshold values; mean: arithmetic mean in mg/l; min: minimum value in mg/l; max: maximum value in mg/l; std: standard deviation of annual concentrations; median: median value of annual concentrations in mg/l; Q1 and Q3, first and third quartile of annual concentrations in mg/l.

Figure captions

Fig. 1. Spatial location of the area of study. (a) Regional scale: Catalonia. (b) Local scale: Camp de Tarragona.

Fig. 2. Groundwater quality control locations in Catalonia.

Fig. 3. Smooth cumulative exposure map for Camp de Tarragona. Combined effect of water pollutants (Pb, Fe, Mn, Ba and Nitrates) exceeding regulatory thresholds for year 2002. The numbers in the labels indicate the number of pollutants exceeding legal threshold values.

Fig. 4. DRASTIC vulnerability map for the Camp de Tarragona.

Fig. 5. SOM-based intrinsic vulnerability model with DRASTIC variables at the local scale. U-matrix and c-planes for the seven DRASTIC features (D, depth to water; R, net recharge; A, aquifer media; S, soil media; T, topography; I, impact of vadose zone; C, hydraulic conductivity)(left); intrinsic vulnerability map for the Camp de Tarragona(right).

Fig. 6. SOM-based specific vulnerability model at the local scale. U-matrix and c-planes for the seven input variables considered in the current specific vulnerability SOM model (H, piezometric level; P, annual rainfall; Ks, soil permeability; %s, land surface slopes; Ka, hydraulic conductivity of the aquifer; land, land use)(left); specific vulnerability map for the Camp de Tarragona(right).

Fig. 7. Spatial distribution of nitrate concentrations for year 2002 generated by: (a) kriging interpolation; and (b) SOM interpolation.

Fig. 8. Cumulative exposure map for Catalonia in year 2002 generated by: (a) kriging interpolation; and (b) SOM interpolation. Combined effect of water pollutants exceeding regulatory thresholds (Pb, Fe, Mn, Se, sulfate and nitrate).The numbers in the labels indicate the number of pollutants exceeding legal threshold values.

Fig. 9. DRASTIC vulnerability map for Catalonia.

Fig. 10. SOM-based specific vulnerability map for Catalonia. The variables used to characterize vulnerability are piezometrics, annual rainfall, soil's permeability, surface slopes, aquifer media, hydraulic conductivity, and land uses.

Fig. 11. Zoom of SOM-based specific vulnerability map for Catalonia to the Camp de Tarragona scale (left); Original scale of the Camp de Tarragona SOM-based specific vulnerability map (right).

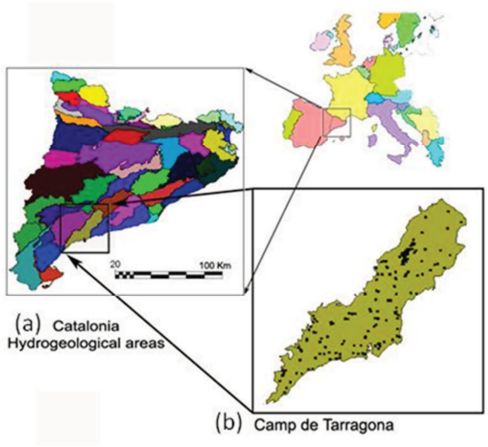


Fig.1

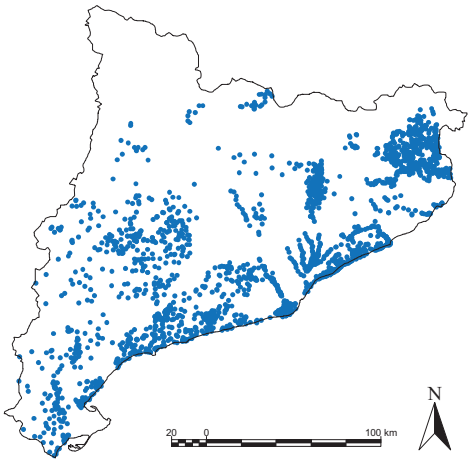


Fig. 2

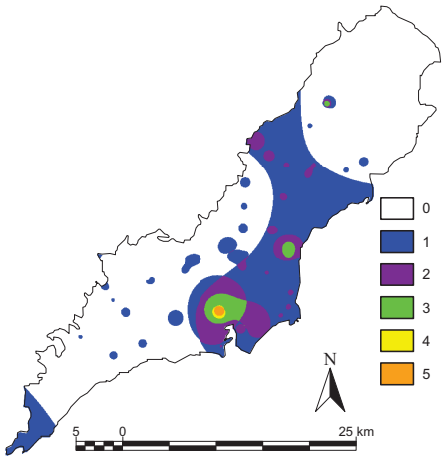


Fig. 3

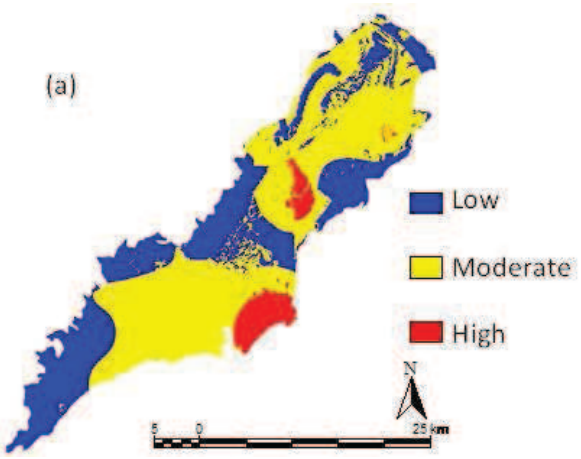
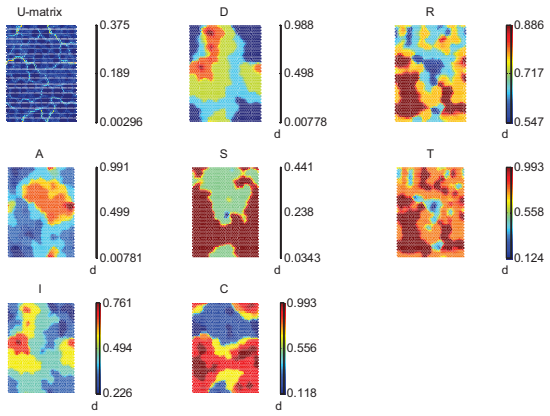


Fig. 4



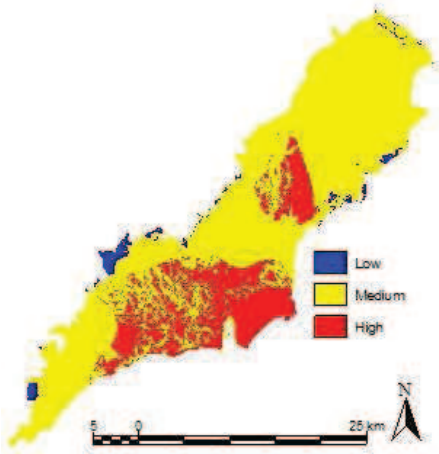
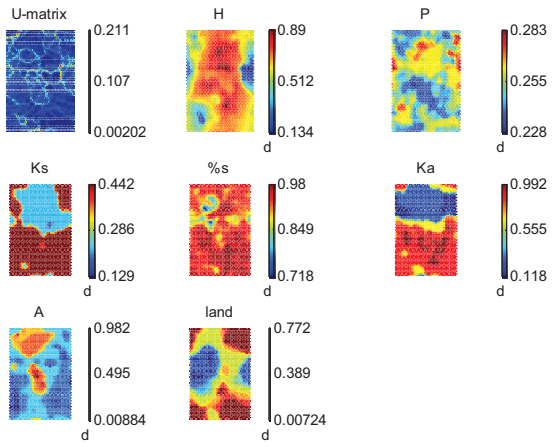


Fig. 5



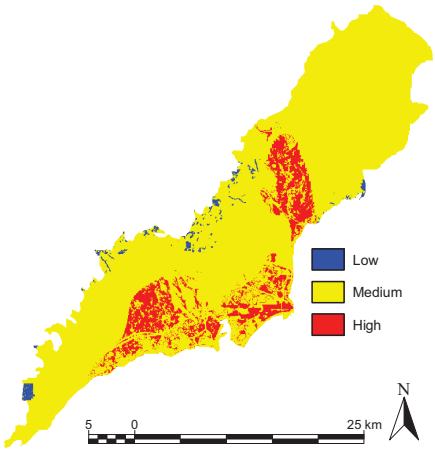


Fig. 6

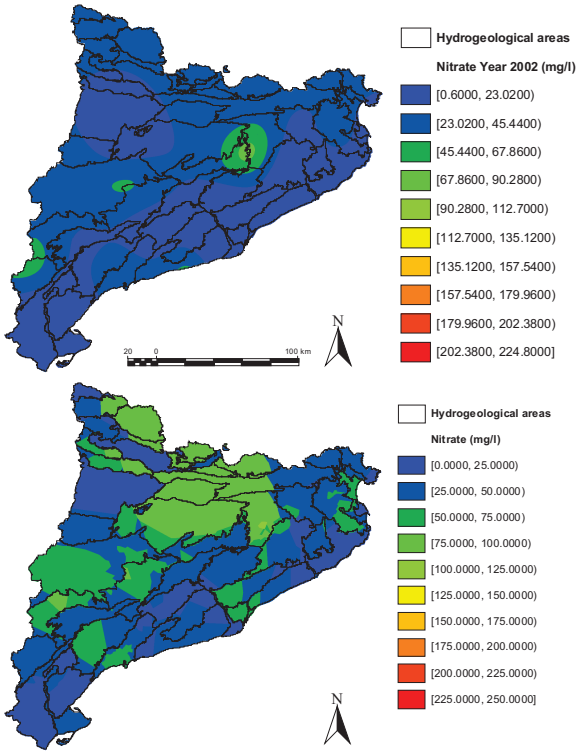

Hydrogeological areas
Nitrate Year 2002 (mg/l)
[0.6000, 23.0200)
[23.0200, 45.4400)
[45.4400, 67.8600)
[67.8600, 90.2800)
[90.2800, 112.7000)
[112.7000, 135.1200)
[135.1200, 157.5400)
[157.5400, 179.9600)
[179.9600, 202.3800)
[202.3800, 224.8000]
20 0 100 km
N
Hydrogeological areas
Nitrate (mg/l)
[0.0000, 25.0000)
[25.0000, 50.0000)
[50.0000, 75.0000)
[75.0000, 100.0000)
[100.0000, 125.0000)
[125.0000, 150.0000)
[150.0000, 175.0000)
[175.0000, 200.0000)
[200.0000, 225.0000)
[225.0000, 250.0000]
N

Fig. 7

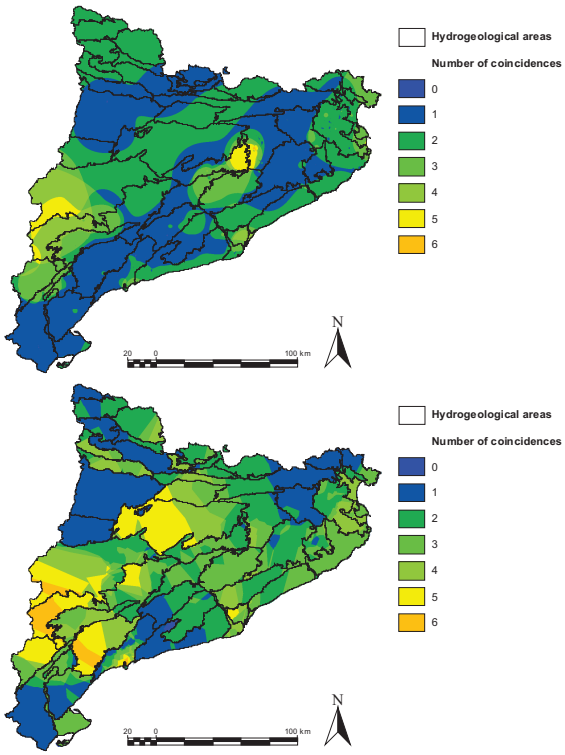

Hydrogeological areas
Number of coincidences
0
1
2
3
4
5
6
20 0 100 km
N
Hydrogeological areas
Number of coincidences
0
1
2
3
4
5
6
20 0 100 km
N

Fig. 8

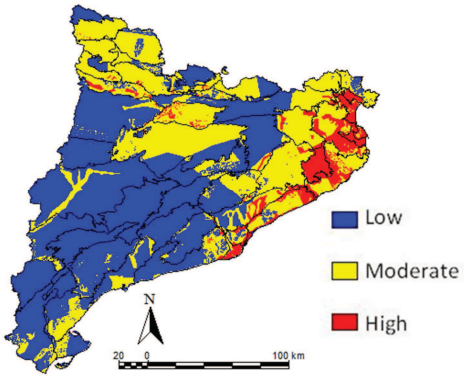


Fig. 9

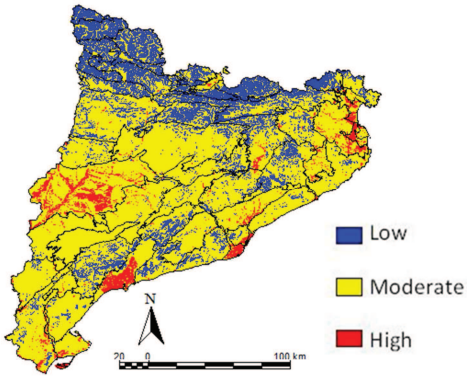


Fig. 10

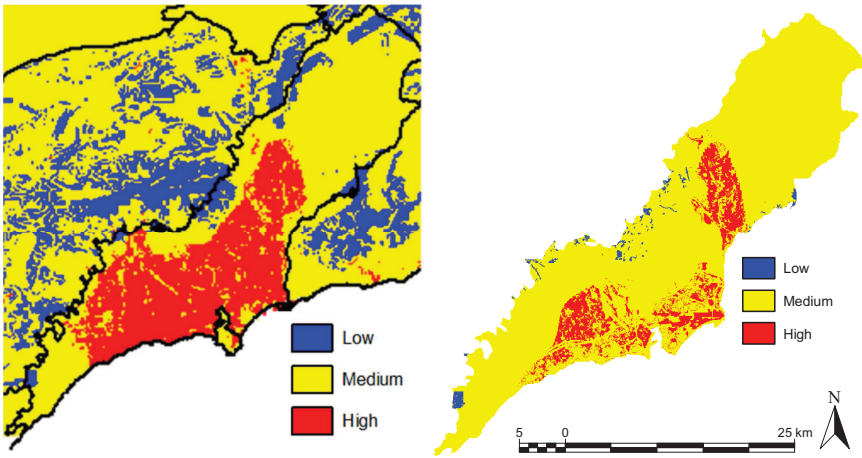


Fig.11

A. Supporting information

Table A.1.
Regulatory limits for pollutants in drinking water

Pollutant	Max. Threshold	Source
NO_2 (mg/l)	50	(a) (b) (c)
NO_3 (mg/l)	0.5	(a) (b)
SO_4 (mg/l)	250	(a) (b)
Fe (µg/l)	200	(a) (b)
Mn (µg/l)	50	(a) (b) (c)
Al (µg/l)	200	(a) (b) (c)
Sb (µg/l)	5	(a) (b) (c)
Ar (µg/l)	10	(a) (b) (c)
Ba (µg/l)	300	(c)
Cd (µg/l)	5	(a) (b)
Cu (µg/l)	2000	(a) (b) (c)
Cr (µg/l)	50	(a) (b) (c)
Mo (µg/l)	0.07	(c)
Ni (µg/l)	50	(a) (b)
Pb (µg/l)	10	(a) (b) (c)
Se (µg/l)	10	(a) (b) (c)
Zn (µg/l)	3000	(c)

Spanish regulatory limit: Real Decreto 140/2003
UE regulatory limit: Council Directive 98/83/EC
World Health Organization: Drinking water standards 1993

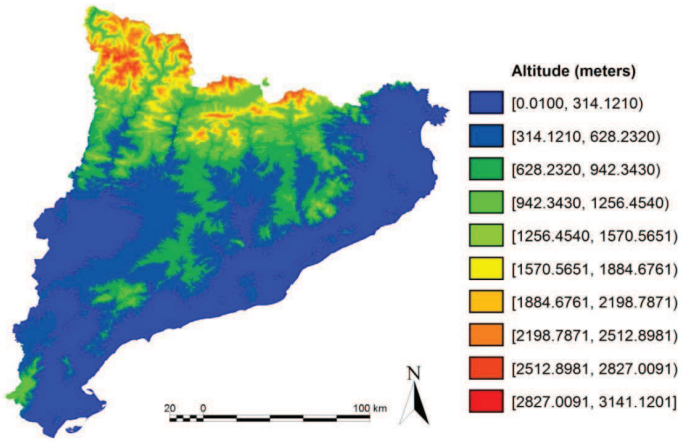


Fig.A.1. Catalonia Digital Terrain Model

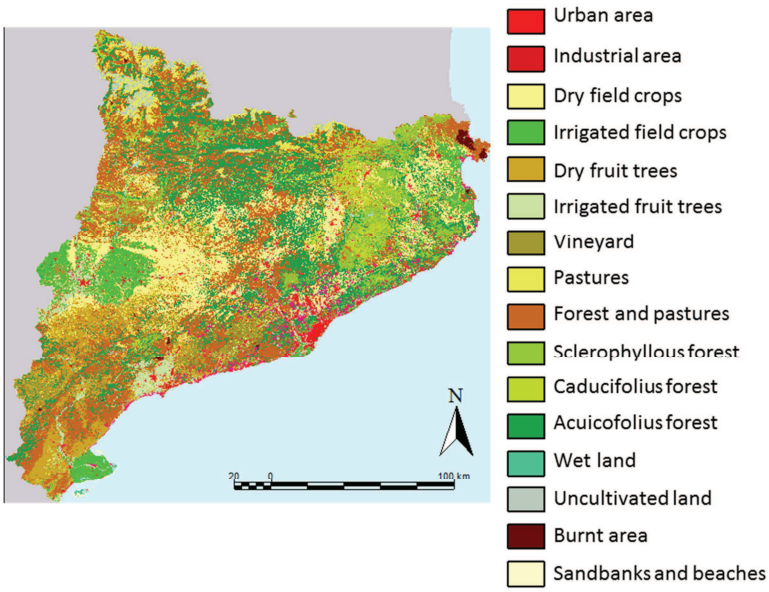


Fig. A.2. Catalonia Land uses map

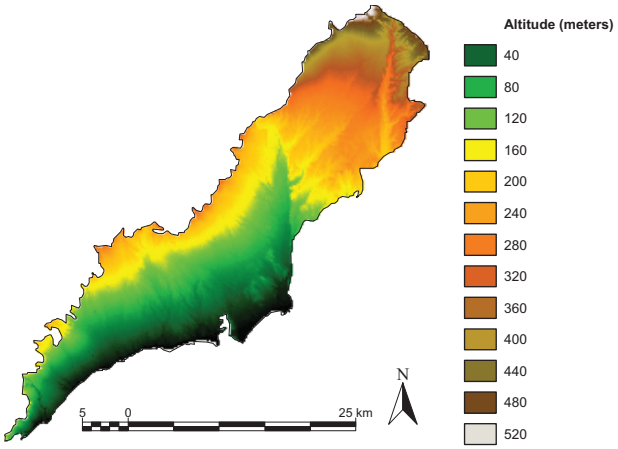


Fig.A.3. Camp de Tarragona Digital Terrain Model

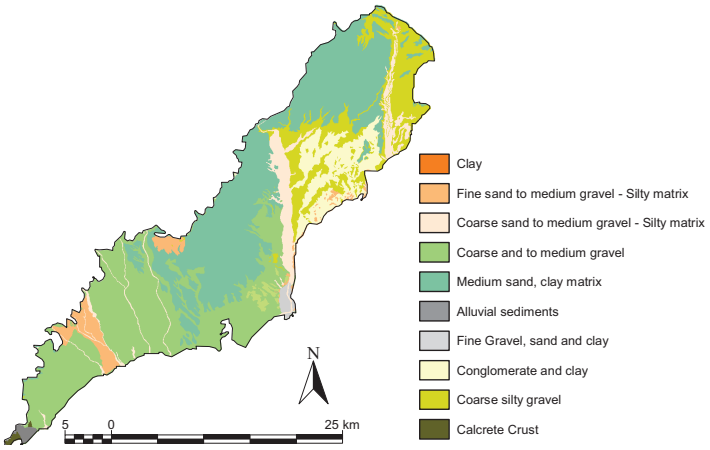


Fig.A.4. Camp de Tarragona Geological Map

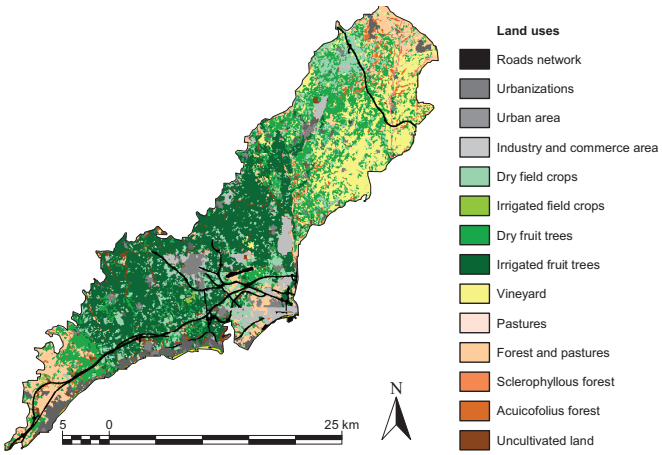


Fig.A.5. Camp de Tarragona Land uses map

B. DRASTIC-based Intrinsic Vulnerability Map for Camp de Tarragona

In order to evaluate DRASTIC index and generate vulnerability maps each parameter in equation (1) has been ranked in a 1-10 rating according to expert criteria (Piscopo, 2001). Table 1 presents the available hydrogeological and climate data, type of data and resolution.

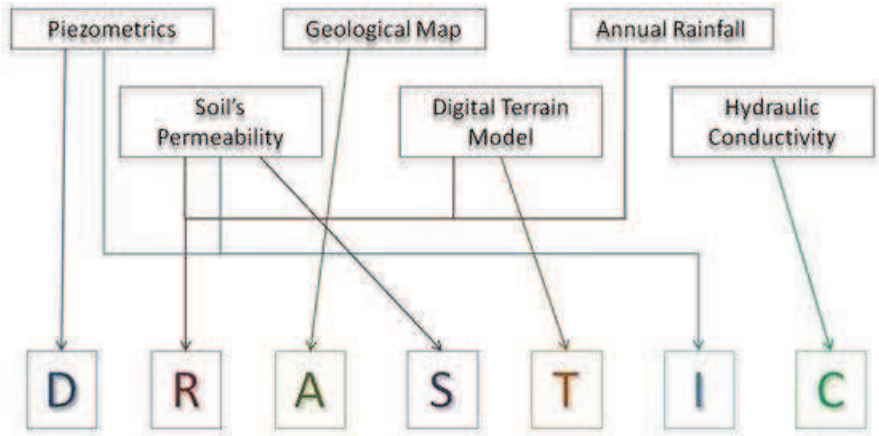


Fig.B.1. DRASTIC features generation from hydrogeological and climate data

Fig.B.1 summarizes the process for generating each DRASTIC feature from available data in the region of study. The DRASTIC index labeling of vulnerability areas was evaluated using different approaches published in the literature. Three labeling approaches were selected. The original one (Aller et al., 1987), that proposed by Draoui et al. (2008) to present a comparative study of vulnerability mapping methods in a Mediterranean area, and the approach of Ahmed (2009) to highlight the differences between Generic and Pesticide DRASTIC models.

B.1.1 Depth to water (D)

Depth to water layer was generated from piezometrics data obtained from the Catalan Water Agency (ACA) and Confederación Hidrológica del Ebro (CHEBRO). Rating categories are summarized in Table B.1 and Fig.B.2.

Table B.1.
Depth to water rating for DRASTIC Index

Range (m)	Rating
0 – 5	10
5 –10	8
10 – 15	6
15 – 20	4
>20	1

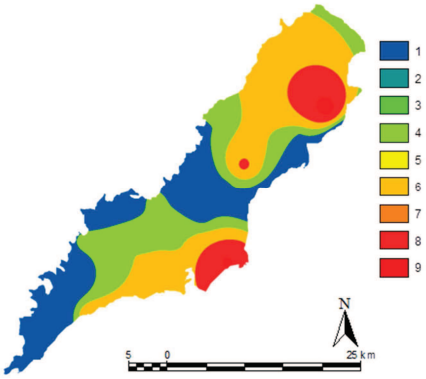


Fig.B.2. Depth to water layer for DRASTIC Index

B.1.2 Net recharge (R)

Based on data gathered from the Generalitat de Catalunya (GENCAT), net recharge parameter has been calculated by a linear combination of annual rain-fall, terrain's slope and soil's permeability (Piscopo, 2001).

Evapotranspiration has been reported constant within the whole Camp the Tarragona area. Equation B.1 presents the calculation of Net Recharge feature, where %s, P and Ks are terrain's slope, annual rainfall and soil's permeability, respectively. They are rated as indicated in Table B.2 and rating categories for Net Recharge are shown in Table B.3and Figure B.3.

$$R = \%s + P + Ks \quad \text{(B.1)}$$

Table B.2.
Ratings for R calculation using equation B.1

%s (%)	Factor	P (mm)	Factor	Ks	Factor
<2	4	>850	4	High	5
2-10	3	700-850	3	Mod-High	4
10-33	2	500-700	2	Mod	3
>33	1	<500	1	Slow	2
				Very slow	1

Table B.3.
Net Recharge rating for DRASTIC Index

Range	Rating
11-13	10
9-11	8
7-9	5
5-7	3
3-5	1

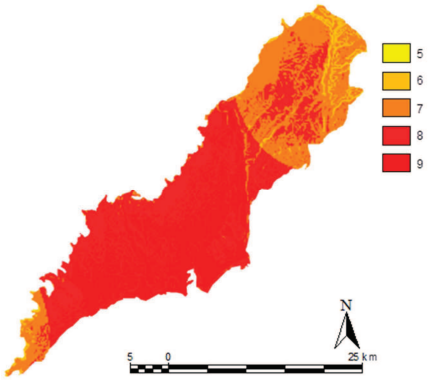


Fig. B.3. Net Recharge layer for DRASTIC Index

B.1.3 Aquifer media (A)

The geological map from GENCAT was processed to reduce the lithology categories. The ratting process has been performed using expert geological criteria. The lithology rating used for the DRASTIC index calculation is presented in Table B.4. Aquifer media layer for Camp de Tarragona area is shown in Figure B.4.

Table B.4.
Aquifer Media ratting for DRASTIC Index

Lithology type	Rating
Calcrete Crust	10
Coarse silty gravel	9
Conglomerate and clay	8
Fine Gravel, sand and clay	7
Alluvial sediments	6
Medium sand, clay matrix	5
Coarse sand to medium gravel	4
Coarse sand to medium gravel- Silty matrix	3
Fine sand to medium gravel - Silty matrix	2
Clay	1

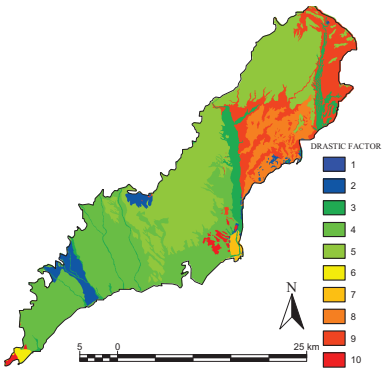


Fig.B.4. Aquifer Media layer for DRASTIC Index

B.1.4 Soil media (S)

The soil media map for the Camp de Tarragona was obtained from the soil's permeability map generated for Catalonia by kriging interpolation of soil's permeability point data provided by CIEMAT.

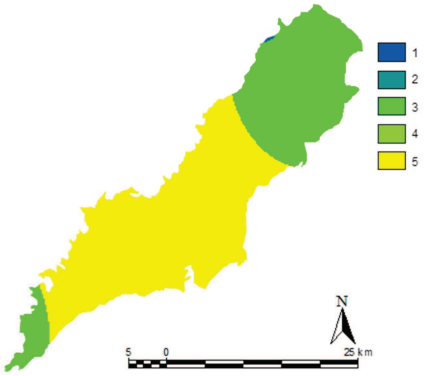


Fig.B.5. Soil Media layer for DRASTIC Index

B.1.5 Topography (T)

Slopes have been calculated from the Digital Terrain Model obtained from Institut Cartografic de Catalunya (ICC). Ratting has been performed based on the literature and expert criteria (Piscopo, 2001)(see Table B.5).

Table B.5.
Topography ratting for DRASTIC Index

Range (%)	Rating
< 2	10
2 – 10	8
10 – 20	5
20 – 33	2
> 33	1

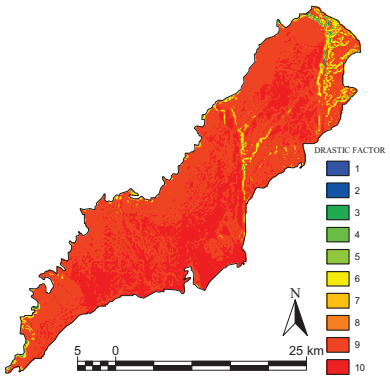


Fig.B.6. Topography layer for DRASTIC Index

B.1.6 Impact of vadose zone (I)

The type of vadose zone determines the attenuation characteristics of the material including the typical soil horizon and rock above the water table. The factors considered important in defining this parameter are soil permeability (or hydraulic conductivity) and depth to water table (Piscopo, 2001). The vadose zone impact is calculated by the following linear additive relation,

Vadose Zone =Soil Permeability + Depth to water table (B.2)

Table B.6.
Factors for Vadose Zone estimation

Soil Permeability	Factor	Depth to water (m)	Factor
High	5	0 – 5	5
Medium High	4	5 –10	4
Medium	3	10 – 15	3
Low	2	15 – 20	2
Very Low	1	>20	1

Table B.7.
Impact of vadose zone ratting for DRASTIC Index

Range	Rating
8 – 10	10
6 - 8	8
4 - 6	5
3 – 4	3
2 - 3	1

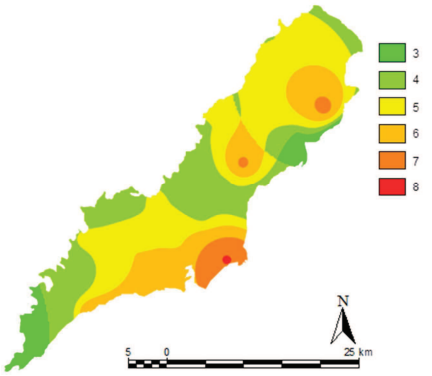


Fig. B.7. Impact of vadose zone ratting for DRASTIC Index

B.1.7 Hydraulic conductivity (C)
This parameter has been obtained from Instituto Geológico y Minero de España (IGME) from a raster map of 1000 x 1000 meters resolution for Spain.

Table B.8.
Hydraulic Conductivity rating for DRASTIC Index

Description	Rating
Very high conductivity – Not consolidated aquifers	10
High conductivity - Not consolidated aquifers	9
Very high conductivity – Karst aquifers	8
High conductivity – Karst aquifers	7
Low conductivity	3
Very low conductivity	2

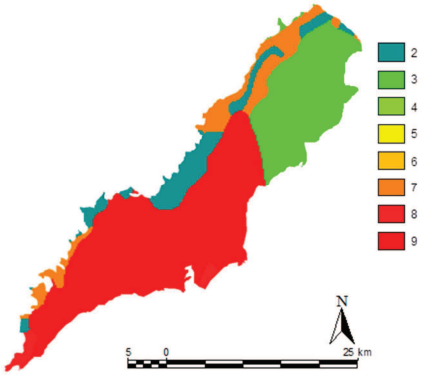


Fig.B.8. Hydraulic Conductivity layer for DRASTIC Index

Annex B

Matlab's Toolboxes Description

B.1 SOM Toolbox in Matlab

Self-organizing maps (SOM) training were performed using Matlab's SOM Toolbox (Vesanto J., J. Himberg, E. Alhoniemi and J. Parhankangas. SOM Toolbox for Matlab. http://www.cis.hut.fi/somtoolbox). Principal functions are presented in Table B.1:

Table B.1. MatLab's SOM Toolbox basic functions

Function	Description
som_normalize	Normalize data set
som_make	Create, initialize and train a SOM
som_quality	Quantization and topographic error of SOM
som_bmus	Calculates best matching units for given data vectors
som_show	Basic visualization of a trained SOM
som_cplane	Calculates component planes
som_umat	Calculates U-matrix

A basic code to train a SOM in Matlab is presented below:

```
%============================================================
% GEOGRAPHIC INTERPOLATION USING SOM
%============================================================


%Initial set-up and cleaning
sD=sDinit;
% Only one component is retained as target for ordinary kriging
sD=som_modify_dataset(sDinit,'extractcomp',desiredData);
[ncases,nvars]=size(sD.data);
target=nvars; % The target variable is located in the last column

%-------------------------------------------------------------
% Detection and filter of outliers in the target value
%-------------------------------------------------------------
 sigma_threshold=3;
m=mean(sD.data(find(~isnan(sD.data(:,target))),target));
s=std(sD.data(find(~isnan(sD.data(:,target))),target));
outliers=find((sD.data(:,target)>m+sigma_threshold*s)|(sD.data(:,target)<m-
sigma_threshold*s));
sD=som_modify_dataset(sD,'removesamp',[outliers]);
sDraw=sD;
```

```
[ncases,nvars]=size(sD.data);

%------------------------------------------------------------
%Normalization of  the UTM coordinates
%------------------------------------------------------------
sD=som_normalize(sDraw,'range');
x=sDraw.data(:,1);
y=sDraw.data(:,2);
sD=som_normalize(sD,'range');

%------------------------------------------------------------
%Training the map
%------------------------------------------------------------
notmissing=find(~isnan(sD.data(:,target)));
myMask=ones(nvars,1); %Define mask in the training process
sM=som_make(sD.data(notmissing,:),'shape','toroid','mapsize','big','training','long',...
        'neigh','cutgauss','mask',myMask,'tracking',0);
[munits,dummy]=size(sM.codebook);
disp(['munits= ' num2str(munits)]);

[qe,te]=som_quality(sM,sD); % Calculate errors in trained SOM
disp(['qe= ' num2str(qe)]);
disp(['te= ' num2str(te)]);

%------------------------------------------------------------
%SOM visualization
%------------------------------------------------------------
sM.comp_names=sD.comp_names;
figure
som_show(sM);

[Pd,V,me,l] = pcaproj(sD,2);
Pm = pcaproj(sM,V,me); % Principal components projection
Code = som_colorcode(Pm); % color coding
hits = som_hits(sM,sD);  % hits
U = som_umat(sM); % U-matrix
Dm = U(1:2:size(U,1),1:2:size(U,2)); % distance matrix
Um=Dm;
Dm = 1-Dm(:)/max(Dm(:)); Dm(find(hits==0)) = 0; % clustering info

bmus = som_bmus(sM,sD); % Best matching units calculation for trained SOM
```

B.2 BME Toolbox in Matlab

Bayesiam maximum entropy (BME) calculations were performed using BMElib Toolbox of Matlab (Bogaert P. and M. Serre. BMElib for Matlab. www.unc.edu/depts/case/BMELIB). The test case for PM_{10} spatio-temporal interpolation for California presented in Christakos (2000) was adapted to the area under study (Catalonia) and the final code used in Matlab is presented below.

```
%============================================================
% SPATIO-TEMPORAL INTERPOLATION USING BME
%============================================================

function
BME_logPM10_catalunya(tMEgrid,recalculate,spaceGrid,contMethod,showDataPoints,...
  showEstPoints,showCounties,showCaBound);

%
% Calculates BME estimates of the annual geometric mean of PM10 at points on
% a spatial grid for selected years, and then plots the corresponding map.
% The data set used are soft probabilistic data for Y(s,t), the annual average
% of log(PM10), available at 191 monitoring stations and for 11 years, from
% 1987 to 1997. The value Y is defined as the annual average of log(PM10),
% where PM10 is the concentration in the air of Particulate Matters of diameter
% smaller than 10 micrometer, and expressed in microgram/cubic meters.
% The BME estimation is performed on the residual Y soft data (i.e. Y minus its
% mean trend), then the mean trend is added back to the Y residual estimate.
% Then the maps are constructed using exp(Y) which is the annual geometric mean
% of PM10, expressed in microgram/cubic meters.
%
% SYNTAX :
%
% examplePM10CA(tMEgrid,recalculate,spaceGrid,contMethod,showDataPoints,...
%   showEstPoints,showCounties,showCaBound);
%
% INPUT :
%
% tMEgrid     1 by k   Optional vector of the selected time (year) for which the
%                      variable is estimated at all the spatial grid points
%                      The default value is [1988 1991 1994 1997]
% recalculate scalar   optional integer indicating wether to recalculate estimates
%                      1 to recalculate the BME estimates.
%                      0 to use already calculated estimates.
%                      default value is 0.
% spaceGrid   string   Optional parameter defining the grid of estimation points
```

```
%               'coarse' for coarse spatial grid covering California
%               'medium' for a grid covering Covering with medium resolution
%               'fine' for a grid covering California on a fine resolution
%               The default value is 'medium'
%               Note: if spaceGrid=[], then the default value is used
% contMethod  scalar   optional integer indicating the method to use to plot contours:
%               1 to plot contour lines,
%               2 to plot a map in color (using pcolor)
%               The default is 2.
% showDataPoints
%         scalar   optional integer indicating whether to show the monitoring stations:
%               1 to show the monitoring stations, 0 ow
%               The default is 1.
% showEstPoints
%         scalar   optional integer indicating whether to show the estimation points:
%               1 to show estimation points, 0 ow
%               The default is 0.
% showCounties
%         scalar   optional integer indicating whether to show the county boundaries:
%               1 to show Counties boundaries, 0 ow
%               The default is 0 if contMethod=1 and 1 if contMethod=2.
% showCaBound scalar   optional integer indicating whether to show the state boundary:
%               1 to show California state boundaries, 0 ow
%               The default is 1.
%
% Set preferences
%
if nargin<1, tMEgrid=[2003 2004 2005 2006 2007]; end;
if nargin<2, recalculate=0; end;
if nargin<3, spaceGrid='fine'; end;
if nargin<4, contMethod=2; end;     % plot a color map
if nargin<5, showDataPoints=1; end; % show the monitoring stations
if nargin<6, showEstPoints=0; end;  % show the estimation points
if nargin<7,                    % show counties if contMethod==2
 if contMethod==1
   showCounties=0;
 else
   showCounties=0;
 end
end
if nargin<8, showCaBound=0; end;    % show boundary of california

plotWhat=1;                     % plot the plot estimated value of annual PM10 geom mean

if exist('./BME_PM10_catalunya.mat')~=2
 disp('Annual geometric mean of PM10 in Catalunya. The first time you run this it will
take some time.');
```

```
else
 disp('Annual geometric mean of PM10 in Catalunya.');
end;
if nargin==0
  disp('Press any key to continue, or press Ctrl C to stop here.');
 pause;
end

%
% Check input variables
%
yr=2003:2007;
for i=1:length(tMEgrid)
 if sum(tMEgrid(i)==yr)~=1,
   error('tMEgrid is a vector specifiying year(s) that must be from 2003 to 2007')
 end
end
if recalculate~=0 & recalculate~=1
 error('recalculate must be equal to 0 or 1');
end
if contMethod~=1 & contMethod~=2
 error('contMethod must be equal to 1 or 2');
end

%
%  Set the covariance model and save it
%
covmodel{1}='exponentialC/exponentialC';
covparam{1}=[0.0015 5000 1];
covmodel{2}='exponentialC/exponentialC';
covparam{2}=[0.001 10000 2];
save PM10_cat_covmodel covmodel covparam;

%
% Create the BME maps of the annual geometric mean of PM10
%
for it=1:length(tMEgrid)
 disp(sprintf('\rProcessing year %d',tMEgrid(it)));
 filename=['logPM10_cat_estBME' num2str(tMEgrid(it))];
 if exist(['./' filename '.mat'])~=2
  recalculate=1;
 end;
 if recalculate==1,
  disp(sprintf('Recalculating the map of BME estimates using a %s grid',spaceGrid));
  pause(0.01);
  estZsBME(spaceGrid,tMEgrid(it),filename)
 elseif recalculate==0,
```

```
   disp('Reading file with already calculated BME estimates');
  elseif recalculate==0,
   error('recalculate must be equal to 0 or 1');
  end;
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%

function estZsBME(spaceGrid,tMEgrid,filename)

% estZsBME  - BME estimation for maps of annual PM10 geometric mean (Jan 1, 2001)
%
% Calculates BME estimates of annual PM10 geometric mean at points on a
% spatial grid for selected years, and then plots the corresponding map.
%
% SYNTAX :
%
% estZsBME(spaceGrid,tMEgrid,filename);
%
% INPUT :
%
% spaceGrid   string   Optional parameter defining the grid of estimation points
%                 'coarse' for coarse spatial grid covering California
%                 'medium' for a grid covering Covering with medium resolution
%                 'fine' for a grid covering California on a fine resolution
%                 The default value is 'medium'
%                 Note: if spaceGrid=[], then the default value is used
% tMEgrid     1 by k   Optional vector of the selected time (year) for which the variable
%                 is estimated at all the spatial grid points
%                 The default value is [1997]
% filename    string   Optional name of file where to save the estimated values
%                 The default value is 'examplePM10CAestBME97'
%
% NOTES :
%
% The number of soft data points used in the neighborhood estimation is 3, 4, or 5
% depending on whether the grid is coarse, medium or fine grid, respectively.
%
% The estimated values saved in filename include the following:
% spaceGrid: a cell array containing two cells, with the x and y coord respectively
%   of the estimation points
% tMEgrid: a vector with the times (in years) of estimation
% XkBME: a nk by length(tMEgrid) array of estimated Z values, where
nk=size(spaceGrid,1)
% XkErr: a nk by length(tMEgrid) array of stan. dev. err
%
```

```
%
% Set preferences
%
%nargin=0;
if nargin<1, spaceGrid='medium'; end;
if nargin<2, tMEgrid=[2003]; end;
if nargin<3, filename='logPM10est_catalunya_2003'; end;

plotErr=1;
plotDisBound=0;
plotEstPoints=1;

if isempty(spaceGrid)
 spaceGrid='fine';
end;

disp('Reading probabilistic data for log yearly logPM10 in Catalunya');
%
% Load the soft data for this study.
%
if exist('./soft_data_logPM10.mat')~=2
 [cs,isST,softpdftype,nl,limi,probdens,filetitle]=readProba('soft_data_logPM10.txt');
 save soft_data_logPM10 cs isST softpdftype nl limi probdens filetitle;
else
 load soft_data_logPM10;
end

%
% Load other useful data
%
[val,valname,filetitle]=readGeoEAS('coord_estaciones_PM10.txt'); % monitoring stations
geographical coordinates
cMS=val;
[val,valname,filetitle]=readGeoEAS('anys_monitoreo_PM10.txt'); % monitoring years
tME=val';
cs=cs{1};               % Don't use index (one variable only)
idxNonEmptyData=nl>0;      % Index of non-empty soft data point
                    % Remove empty soft data points
[cs,c2,nl,limi,probdens,nl2,limi2,probdens2]=probasplit(cs,...
 softpdftype,nl,limi,probdens,idxNonEmptyData);

%
% Select the grid at which estimation will be done
%
disp('Setting parameters');
load catalunya_grids;
```

```
if ischar(spaceGrid)
 switch spaceGrid
 case 'coarse',   %%%%dx=dy=1000
  xk=utmx_coarse_act;
  yk=utmy_coarse_act;
  limite=limiteCat_coarse;
 case 'medium',   %%%%dx=dy=600
  xk=utmx_medium_act;
  yk=utmy_medium_act;
  limite=limiteCat_medium;
 case 'fine',     %%%%dx=dy=200
  xk=utmx_fine_act;
  yk=utmy_fine_act;
  limite=limiteCat_fine;
 otherwise
  error('bad char string for spaceGrid')
 end;
else
 xk=spaceGrid(:,1);
 yk=spaceGrid(:,2);
end;

%
%  Set the covariance model
%
load PM10_cat_covmodel;

%
% set the BME paramaters
%
nhmax=0;              % max number of hard data
switch spaceGrid
case 'coarse',
 nsmax=3;            % max number of soft data
 maxpts=50000;        % number of function eval for integration
case 'medium',
 nsmax=4;            % max number of soft data
 maxpts=100000;       % number of function eval for integration
case 'fine',
 nsmax=5;            % max number of soft data
 maxpts=1000000;      % number of function eval for integration
end;
order=NaN;            % do not estimation the mean trend
dmax=[2000 1.5 5];      % dmax(1)=2000 m, spatial search radius
                % dmax(2)=1.5 yr, temporal search radius
                % dmax(3)=5 m/yr, space/time metric
rEps=0.05;            % Relative numerical error allowable
```

```
nMom=2;              % Calculate the mean and variance
options=BMEoptions;
options(1)=1;
options(3)=maxpts;
options(4)=rEps;
options(8)=nMom;

%
% Get the S/T mean trend mst and adjust the hard and soft data
%
[val,valname,filetitle]=readGeoEAS('logPM10_catalunya_promedios_estaciones.txt');
ms=val;
[val,valname,filetitle]=readGeoEAS('logPM10_catalunya_promedios_anys.txt');
mt=val';
mstgrid=stmeaninterp(cMS,tME,ms,mt,cMS,tME);
[cmst,mst]=valstg2stv(mstgrid,cMS,tME);
mst=mst(idxNonEmptyData);
[limia]=probaoffset(softpdftype,nl,limi,-mst);

XkBMEa=nan*zeros([length(xk) length(tMEgrid)]);
XkErr=nan*zeros([length(xk) length(tMEgrid)]);
XkBME=nan*zeros([length(xk) length(tMEgrid)]);

%
% For each selected year, estimate Z using BMEprobaMoments
%
for i=1:length(tMEgrid),
  tk=tMEgrid(i);
  dateyear=num2str(tk);
  disp(sprintf('Calculating map for year %s',dateyear));
  %
  %  Estimate the Z using BMEprobaMoments
  %
  ck=[xk yk tk+0*xk];
  ch=zeros(0,3);
  zh=zeros(0,1);

[moments,info]=BMEprobaMoments(ck,ch,cs,zh,softpdftype,nl,limia,probdens,covmodel,
covparam,nhmax,nsmax,dmax,order,options);
  XkBMEa(:,i)=moments(:,1);
  XkErr(:,i)=sqrt(moments(:,2));
  %
  % Re-ajust XkBMEa by adding the mean to the estimated values
  %
  XkBMEtk=XkBMEa(:,i);
  XkBMEtk(isnan(XkBMEtk))=0;
  msttk=stmeaninterp(cMS,tME,ms,mt,[xk yk],tk);
```

```
  XkBME(:,i)=XkBMEtk+msttk;
  %
  %  Re-ajust XkErr by setting NaN to max of standard dev error
  %
  XkErrtk=XkErr(:,i);
  XkErrtk(isnan(XkErrtk))=max(XkErrtk);
  if ~isreal(XkErrtk), XkErrtk=real(XkErrtk); end;
  XkErr(:,i)=XkErrtk;
end;
XkBME=exp(XkBME);
prueba=limite;
prueba(limite==0)=-32768;
prueba(limite==1)=XkBME;
XkBME=prueba;
prueba(limite==1)=XkErr;
XkErr=prueba;
%
% Save the output file and plot the map
%
spaceGrid=[xk yk];
save(filename,'spaceGrid','tMEgrid','XkBME','XkErr','cMS','tME');
```