Environmental Risk Assessment in the Tuscany Region: a Proposal

ENVIRONMETRICS
Environmetrics 2007; **18**: 315–332
Published online 19 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.831

Environmental risk assessment in the Tuscany region: a proposal

Giovanna Jona Lasinio[1], Fabio Divino[2]*,† and Annibale Biggeri[3]

[1] *Università di Roma 'La Sapienza', Italy*
[2] *Università del Molise, Italy*
[3] *Università di Firenze, Italy*

SUMMARY

In this paper we present a first attempt to define a statistically coherent protocol for Environmental Risk Assessment (ERA), considered as a classification problem. Our approach moves from an idea developed in the pattern recognition literature. Several independent classifiers, each working on a subset of the covariates space, produce a set of corresponding units classifications. Then a *gate*, modulating the partial results, produces a final, combined classification. We propose a combined classification strategy based on rank transformations and Bayesian mixture classifiers. Although the use of Bayesian mixture models in classification problems is quite common in many fields of application, the novelty of our proposal concerns the use of truncated Gaussian components to model the behaviour of the rank variables in a multidimensional setting. We approach the general problem by partitioning the covariate space into several subspaces, each one representing one environmental dimension; we consider three environmental dimensions (Air, Water and Waste), represented by several pressure indicators. The evaluation of environmental risk for the Tuscany municipalities is our motivating example. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: environmental risk; multiple ranks; Bayesian mixture classifiers; truncated Gaussian components; environmental indicators

1. INTRODUCTION

Hazard is commonly defined as 'the potential to cause harm'. A hazard can be defined as 'a property or situation that in particular circumstances could lead to harm' (Royal Society, 1992). Risk is a more difficult concept to define. The term risk is used in everyday language to mean 'chance of disaster'. When used in the process of risk assessment it has specific definitions, the most commonly accepted being 'the combination of the probability, or frequency, of occurrence of a defined hazard and the magnitude of the consequences of the occurrence' (Royal Society, 1992).

The distinction between hazard and risk can be made clearer by the use of a simple example. A large number of chemicals have hazardous properties. Acids may be corrosive or irritant to human beings for example. The same acid is only a risk to human health if humans are exposed to it. The degree of harm

*Correspondence to: F. Divino, DSPSA University of Rome 'La Sapienza', Rome, Italy.
†E-mail: fabio.divino@uniroma1.it

Received 23 August 2006
Accepted 19 December 2006

caused by the exposure will depend on the specific exposure scenario. If a human only comes into contact with the acid after it has been heavily diluted, the risk of harm will be minimal but the hazardous property of the chemical will remain unchanged.

There has been a gradual move in environmental policy and regulation from hazard- to risk-based approaches. This is partly due to the recognition that for many environmental issues a level of zero risk is unobtainable or simply not necessary for human and environmental protection and that a certain level of risk in a given scenario is deemed ‘acceptable’ after considering the benefits.

In broad terms risk assessments are carried out to examine the effects of an agent on humans (Health Risk Assessment [HRA]) and ecosystems (Ecological Risk Assessment [EcoRA]). Environmental Risk Assessment (ERA) is the examination of risks resulting from technology that threaten ecosystems, animals and people. It includes human health risk assessments, ecological or ecotoxicological risk assessments and specific industrial applications of risk assessment that examine end-points in people, biota or ecosystems.

Many organisations are now actively involved in ERA, developing methodologies and techniques to improve this environmental management tool. Such organisations include the Organisation for Economic Co-operation and Development (OECD), the World Health Organization (WHO) and the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC). One of the major difficulties concerning the use of risk assessment is the availability of data and the data that are available are often loaded with uncertainty.

Risk assessment is carried out to enable a risk management decision to be made. It has been argued that the scientific risk assessment process should be separated from the policy risk management process but it is now widely recognised that this is not possible. The two are intimately linked.

Risk management is the decision-making process through which choices can be made between a range of options which achieve the ‘required outcome’. The ‘required outcome’ may be specified by legislation, by way of environmental standards, may be determined by a formalised risk-cost-benefit analysis or may be determined by another process, for instance ‘industry norms’ or ‘good practice’. It should result in risks being reduced to an ‘acceptable’ level within the constraints of the available resources.

The main aspect of ERA is the multidisciplinary nature of the problem. ERA has traditionally been a function of policy and regulatory agencies and most development has taken place in these fields. Up to now the most established procedures concern HRA, and as this approach lends itself well in many respects to EcoRA, it has been the starting point for EcoRA procedures definitions. However, due to the complex nature of the potential target(s) or receptor(s), several problems have presented themselves to practitioners. HRA is concerned with individuals and morbidity and mortality; EcoRA is concerned with populations and communities and the effects of substances on mortality and fecundity (Fairman *et al.*, 1999). EcoRA is very much a developing field and has many problems which need resolving such as:

- determining the effects at population and community level;
- selection of end-points;
- selection of indicative species;
- selection of field, laboratory, mesocosm and microcosm tests;
- incorporation of resilience and recovery factors of the ecosystem.

These problems are covered comprehensively in such texts as Suter (1993) and Van Leeuwen and Hermens (1995).

In other words a great amount of work has already been done in order to choose what to measure, and how and what to observe to obtain reliable evaluation of human risk and environmental risk leading to

DOI: 10.1002/env

establish ERA techniques adopted by several environmental protection agencies all over the world, a good source of information on this topics is Fairman *et al.* (1999). On the other hand, only recently research focused on statistical techniques capable of improving risk assessment procedures. This fact implies that problems, such as data quality or how to give a joint risk evaluation (from the health and ecological perspectives), are not usually taken into account during the elaborations. Often uncertainty evaluation is not rigorously carried out and very rarely one can find a complete vision of an environmental system fully accounted (see for instance Various Authors, 2003).

In this paper we propose a possible approach to the risk assessment of given spatial units (the Tuscany municipalities in this paper). We approach the problem as a complex classification one. Our aim is to classify Tuscany Municipalities according to their environmental status and potential status. More precisely, referring to the DPSIR (Driving force-Pressure-State-Impact-Response) model, proposed in 1995 by the EEA, we divide available data into Pressure and Status indicators, then we classify municipalities using pressure indicators in order to evaluate the potential environmental risk and according to status indicators in order to evaluate the real situation. This last step will be developed in further work when data will be available.

In principle, in order to give a complete risk evaluation, really useful to risk managers, we should include all kinds of information in our data set: health-related variables, ecological variables, environmental indicators, demographical information, economical data and so on. Then the first problem we have to face is data availability. 'Good' data are usually not available and in several cases it is not clear which type of indicators would be appropriate to include in order to get a 'sensible' joint evaluation.

Assuming that we can collect enough information to start the evaluation process, a second problem arises: dimensionality. In classical classification problems we should consider all variables/indicators simultaneously in order to build the classification; this will not be computationally feasible with standard techniques. Our proposal starts from an idea born in the pattern recognition literature, in which the combination of several classifiers, each working on a subset of the covariates space, produces a set of unit classifications and then a *gate* produces the final, combined classification. The general combined scheme is illustrated in Figure 1.

Mixture of experts (ME) and their hierarchical version (HME) were introduced by Jacobs *et al.* (1991) and by Jordan and Jacobs (1994), respectively. Their general aim was to find an automatic learning technique that could present a different structure (according to a given criterion) in different regions of the covariates space. For instance, in the ME architecture, experts are usually feed-forward multilayer neural networks or one-layer structures (for instance generalised linear models). These neural networks are trained simultaneously in such a way that each expert competes with the others in learning the input patterns and becoming 'responsible' for a specific region of the covariate space. The competition is overlooked by the *gating network* which builds a sequence of weights evaluating the single expert performance on each region. This sequence of weights depends on the input variables. Here we refer to a generalisation of ME and HME with respect to the type of architecture used in the experts combination and input structure, and it has been already implemented in problems of speaker recognition (Brutti *et al.*, 2002; Jona Lasinio *et al.*, 2004).

In order to establish the formal representation of this approach, let us illustrate in detail how the simplest architecture (ME) works. Each expert can be seen as a statistical process generating an output $\mathbf{y}$ according to a given probability function $g(\mathbf{y}|\mathbf{x},\boldsymbol{\xi})$ depending on the input variable $\mathbf{x}$ and a parameters set $\boldsymbol{\xi}$. The gating network can be represented as a classifier (a parametric model) that, conditional on a parameters vector $\boldsymbol{\gamma}$, transforms the covariate vector $\mathbf{x}$ into a probability vector $\{w_j(\mathbf{x},\boldsymbol{\gamma})\}_{j=1,\ldots,N(e)}=\{\mathbf{P}(e_j|\mathbf{x},\boldsymbol{\gamma})\}_{j=1,\ldots,N(e)}$ ($N(e)$ being the number of experts and e_j being the j-th

DOI: 10.1002/env

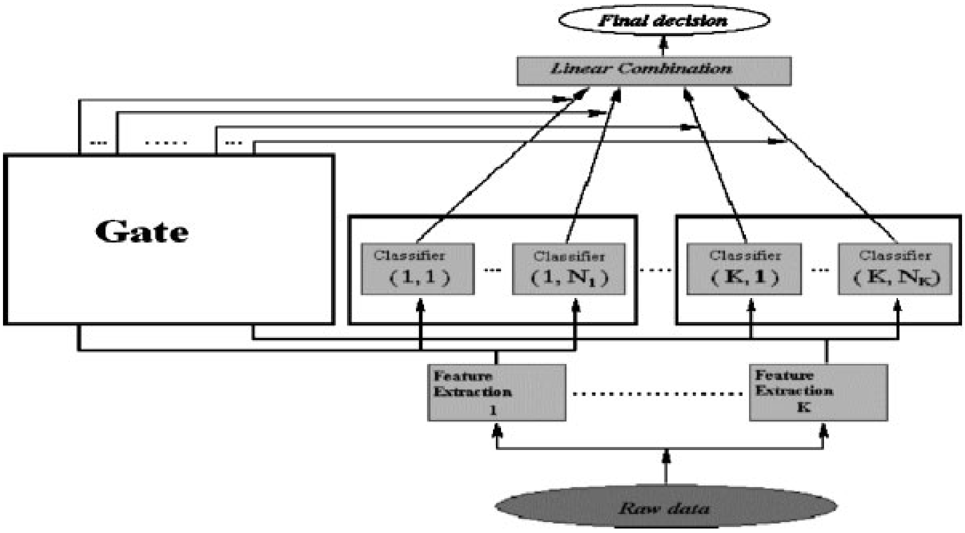

Figure 1. Combined classification strategy

expert) which reports the probability for each expert to return the 'correct classification'. Then the gating network in assigning the probability vector $\mathbf{w}(\mathbf{x},\boldsymbol{\gamma}) = [w_1(\mathbf{x},\boldsymbol{\gamma}),\ldots,w_{N(e)}(\mathbf{x},\boldsymbol{\gamma})]^T$ to a point in the covariate space X generates a fuzzy partition of X into $N(e)$ regions $\{\mathbf{R}_1,\ldots,\mathbf{R}_{N(e)}\}$. Each $\mathbf{R}_\mathrm{j}$ is controlled by the expert e_j through the function $w_j(\mathbf{x},\boldsymbol{\gamma})$. The probability of observing a correct behaviour of the entire ME model is given by the following mixture (McLachlan and Peel, 2000):

$$g(\mathbf{y}|\mathbf{x},\boldsymbol{\xi}) = \sum_{j=1}^{N(e)} w_j(\mathbf{x},\boldsymbol{\gamma})g(\mathbf{y}|\mathbf{x},\xi_j)$$

Once parameters $\{\xi_j\}_{j=1,\ldots,N(e)}$ and $\boldsymbol{\gamma}$ are estimated we can synthesise the above distribution by its mean vector. The above architecture is completely general, we can specify any type of model for the experts and gating. In this paper a HME approach will be adopted.

In ERA studies this architecture must be adapted to the huge complexity of the problem. For instance the number of classes must be defined according to environmental protection regulation. Missing data, misaligned data, low-quality data and so on must be properly treated in order to reduce their influence on the general classification mechanism. However, this approach seems quite feasible at least from a logical point of view as it allows to drastically reduce the dimensionality of the problem. Furthermore in ERA procedure a natural partition of the covariate space can be easily found. We can think of our problem as made of smaller subproblems, each one dealing with one aspect of the Environment: Air, Water, Soil, Human Health and so on.

In this paper we present a first attempt to the implementation of a combined strategy in which a mixture-based classifier (Section 2) combined with 'researcher intervention' (see Sections 5 and 6 for detail) is used as *gate* in the general architecture. One advantage of the proposed approach is the semi-automatic nature of the classification process, here we do not have to fix *a priori* the number of classes, we can rely on 'objective' statistical criteria to choose it (see Section 5).

DOI: 10.1002/env

The paper is organised as follows. In Section 2 we present a Bayesian mixture classification model. In Section 3 we report a detailed description of available data, followed, in Section 4, by the exploratory data analysis (EDA) and the subspaces classifications. In Section 5 we illustrate the final classification results and in Section 6 we draw conclusions pointing to several further developments.

2. BAYESIAN MIXTURE CLASSIFICATION MODEL

When subspaces partial classifications are obtained we need to combine them in a unique output. Furthermore we want this final units ordering to be of use for environmental decision makers. With this in mind, we present a classification technique based on a Bayesian parametric mixture model.

The use of Bayesian mixture in classification problems is quite common in many fields of application, see for example Green and Richardson (2002) for an application in disease mapping. Here we propose a Bayesian mixture model to classify multi-rank variables; in particular we will introduce the use of truncated Gaussian mixture components to model the behaviour of the rank transformations.

In order to define a general framework, we assume that data to be classified are sampled from a mixture of a fixed number G of components $f(y;\theta_1),\ldots,f(y;\theta_G)$ in different and unknown normalised proportions $\pi_1,\ldots,\pi_G$. The model can be represented as follows:

$$f(\boldsymbol{y};\boldsymbol{\psi}) = \sum_{k=1}^{G} \pi_k f(y;\theta_k)$$

where y is a m-dimensional response variable and $\boldsymbol{\psi} = (\pi_1,\ldots,\pi_G,\theta_1,\ldots,\theta_G)$ is the set of weights and parameters characterising the mixture (McLachlan and Peel, 2000). The probability function of a random sample $y = (y_1,\ldots,y_n)$ drawn from the above model is given by:

$$f(\boldsymbol{y};\boldsymbol{\psi}) = \prod_{i=1}^{n}\sum_{k=1}^{G} \pi_k f(y_i;\theta_k)$$

The previous expression presents analytical problems. In order to manage this difficulty, we can consider a missing data structure and associate to each observation i, a latent label $Z_i \in \{1,2,\ldots,G\}$ indicating which component of the mixture has generated the related observed value y_i. In this way, each weight π_k represents the probability that the label Z_i is represented by k.

In practice, it is convenient to deal with a G-dimensional vector $\mathbf{Z}_i = (Z_{i1},\ldots,Z_{iG})$ instead of the categorical label Z_i. Then the generic k element Z_{ik} is defined to be one or zero whether the generating component of y_i is k-th or not. The point can be better formalised by assuming that each vector $\mathbf{Z}_i$ is distributed as a multinomial random variable consisting of only one drawn over G categories with probabilities given, respectively, by the weights $\pi_1,\ldots,\pi_G$ (McLachlan and Peel, 2000). Then, the complete data distribution has the following form:

$$f(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\psi}) = \prod_{i=1}^{n}\prod_{k=1}^{G} [\pi_k f(y_i;\theta_k)]^{z_{ik}}$$

This kind of approach seems to have a natural framework in the Bayesian setting, in fact the weights can be considered as prior guesses that each observation belongs to the several latent groups. In order to

DOI: 10.1002/env

complete the specification of the model, we need to introduce also a prior distribution over the hyperparameter $\boldsymbol{\psi}$ that we denote by $P(\boldsymbol{\psi})$. Then, in order to make inference on the quantities of interest, we can consider the full posterior distribution:

$$P(\boldsymbol{\psi}, z|y) \propto P(\boldsymbol{\psi}) \prod_{i=1}^{n} \prod_{k=1}^{G} [\pi_k f(y_i; \theta_k)]^{z_{ik}}$$

Estimates of weights $\hat{\pi}_k$ and parameters $\hat{\theta}_k$ can be computed by simulating the marginal posterior profiles through Markov chain Monte Carlo techniques (Robert and Casella, 1999).

In data classification the main purpose is to infer the latent labels $Z_1, \ldots, Z_n$ on the basis of the responses $y_1, \ldots, y_n$. After a mixture model has been fitted, we can produce a probabilistic classification of those observations in terms of their posterior probabilities of belonging to the different classes. In fact for each observation we can compute the quantities:

$$\tau_k(y_i; \hat{\boldsymbol{\psi}}) = \frac{\hat{\pi}_k f(y_i, \hat{\theta}_k)}{\sum_{h=1}^{G} \hat{\pi}_h f(y_i, \hat{\theta}_h)} \quad k = 1, \ldots, G$$

which represent the estimated posterior probabilities that the i-th observation has been drawn, respectively, from the first, the second, ..., the G-th latent group. Then, we can classify the data by assigning each y_i to the class to which it has the highest estimated posterior probability of belonging. In other words we estimate the allocation label by setting the components of the related vector $\boldsymbol{Z}_i$ as:

$$\hat{Z}_{ik} = \begin{cases} 1, \text{ if } k = \arg - \max\limits_{h=1,\ldots,G} \tau_h(y_i, \hat{\boldsymbol{\psi}}) \\ 0, \text{ otherwise} \end{cases}$$

that produces the optimal empirical Bayesian classification (McLachlan and Peel, 2000).

2.1. *Truncated discrete Gaussian mixture for multi-rank variables*

In many empirical applications it is convenient to deal with the rank transformation of the observed responses. This is the case when a set of several features measured or estimated over heterogeneous scales are considered, and it becomes more important to preserve the robustness than the precision of the adopted methodologies. With this in mind, here we consider that the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ to be classified are represented by multivariate rank variables, that is $y_i = (y_{i1}, \ldots, y_{iD})$ where each $y_{id} \in \{1, \ldots, n\}$ for $d = 1, \ldots, D$. More precisely, in our application, the rank variables represent the output of the chosen classifier operating on each covariates subspace.

We suppose that observations drawn from the same mixture component will present a homogeneous pattern with respect to the several dimensions $d = 1, \ldots, D$. Under the hypothesis of independence, this point can be formalised by assuming that each component of the mixture is the product of D truncated discrete Gaussian distributions:

$$(y_i; \theta_k) \sim \prod_{d=1}^{D} \frac{\exp\left\{-\frac{S_{kd}}{2}(y_{id} - \mu_{kd})^2\right\}}{\sum_{r_d=1}^{n} \exp\left\{-\frac{S_{kd}}{2}(y_{id} - \mu_{kd})^2\right\}} \tag{1}$$

In this way we can enclose into the mixture model the homogeneity of the Gaussian approach, while preserving the discrete nature of the data.

DOI: 10.1002/env

As we mentioned in the previous section, in order to complete the Bayesian scheme, we need to introduce prior distributions over all quantities characterising the mixture. They are the mean parameters μ_{kd} and the precision parameters S_{kd}. More precisely we suppose that the mean of each component k, with respect to each dimension d, is uniformly distributed over the continuous range of the ordinal scale, that is:

$$\mu_{kd} \sim \frac{1}{n-1} \quad \text{with } \mu_{kd} \in [1, n]$$

This assumption is coherent with the underlying idea that we are classifying data observed over several independent dimensions and we cannot assume any particular *a priori* average behaviour of the rank variables. With respect to each precision parameter S_{kd}, we adopt a standard choice by defining a Gamma distribution over its range:

$$S_{kd} \sim \frac{\beta_{kd}^{\alpha_{kd}}}{\Gamma(\alpha_{kd})} \mathrm{e}^{\beta_{kd} S_{kd}} S_{kd}^{\alpha_{kd}-1} \quad \text{with } S_{kd} > 0,\ \alpha_{kd} > 0,\ \beta_{kd} > 0$$

where the hyperparameters α_{kd} and β_{kd} are fixed *a priori*. In general, a proper strategy to choose those hyperparameters is to consider that each prior distribution should be quite flat over the range (noninformative strategy). In order to preserve this idea, after an empirical tuning over a grid of several values had been performed, we fixed each $\alpha_{kd} = 1.01$ and each $\beta_{kd} = 0.01$. For the normalised weights vector $\pi = (\pi_1, \ldots, \pi_G)$ we choose a Dirichlet prior distribution:

$$\pi \sim \Gamma\left(\sum_{k=1}^{G} \delta_k - G\right) \prod_{k=1}^{G} \frac{\pi_k^{\delta_k - 1}}{\Gamma(\delta_k)} \quad \text{with } \pi_k \in [0, 1],\ \delta_k > 0$$

that represents the conjugate prior with respect to mixture components belonging to the exponential family (McLachlan and Peel, 2000). As a noninformative prior guess, we fix each $\delta_k = 1.01$.

Under these assumptions we can derive the full joint posterior distribution of the labels $\boldsymbol{z}$ and hyperparameter $\boldsymbol{\psi}$, given the rank data $\boldsymbol{y}$, as:

$$\begin{aligned} P(z, \boldsymbol{\psi} | y) \sim \exp\Bigg\{ & \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \left[\log \pi_k - \frac{1}{2} \sum_{d=1}^{D} S_{kd} (y_{id} - \mu_{kd})^2 - \sum_{d=1}^{D} \log \sum_{r_d=1}^{R} \mathrm{e}^{-\frac{S_{kd}}{2}(r - \mu_{kd})^2} \right] \\ & + \sum_{k=1}^{G} (\delta_k - 1) \log \pi_k - \sum_{k=1}^{G} \log \Gamma(\delta_k) + \log \Gamma\left(\sum_{k=1}^{G} \delta_k - G \right) - G \cdot D \log(n-1) \\ & + \sum_{k=1}^{G} \sum_{d=1}^{D} [(\alpha_{kd} - 1) \log S_{kd} - \beta_{kd} S_{kd} + \alpha_{kd} \log \beta_{kd} - \log \Gamma(\alpha_{kd})] \Bigg\} \end{aligned} \qquad (2)$$

This posterior can be simulated via a Gibbs-Metropolis sampler. In order to classify the data we need to compute the posterior probabilities $\tau_k(y_i, \hat{\boldsymbol{\psi}})$ and then apply the optimal empirical Bayes assignment previously introduced.

3. AVAILABLE DATA

Here we describe the data sets that have been made available to us. They refer to three environmental matrices: Water, Waste and Air, and report only pressure indicators. Data are aligned in space but they are misaligned in time.

DOI: 10.1002/env

3.1. Water

Anthropic pressure on water resources is obtained by estimating potential loadings of different nature. The estimation procedure is illustrated in Various Authors (1991), a technical report from Consiglio Nazionale delle Ricerche-Istituto di Ricerca sulle Acque (CNR-IRSA). Here we simply sketch the general technique.

Potential organic loadings are expressed in terms of Inhabitants Equivalent (IE); in other words one unit of loading corresponds to potential pollutant load produced by and individual during 24 h in a year. The total potential organic load is obtained considering several human activities from agriculture to tourism. An example of conversion coefficients is reported below:

- 1 Resident = 1.00 IE
- Tourists/365 = 1.00 IE
- 1 Bovine = 8.16 IE
- 1 Equine = 8.08 IE
- 1 Ovine-Caprine = 1.78 IE
- 1 Swine = 1.95 IE
- 1 Poultry = 0.20 IE

Potential trophic loads estimate the amount of nitrogen and phosphorus introduced in the environment by human activities. It is measured in kilograms and the estimation procedure is analogous to the others here briefly illustrated.

In other words, all pressure indicators are estimates based on specific models described in Various Authors (1991). The territorial reference unit is the municipality, estimates are available by sources (population, industry and agriculture). Our choice is to work with the total estimated loadings of Organic (IE), Nitrogen (kg) and Phosphor (kg). Population estimates are based on the 1998 population and tourist activities Census; industrial estimates are based on the 1996 Census of industrial activities and agriculture estimates comes from data collected during the 2000 Census of agriculture.

3.2. Air

The Air matrix, we are going to analyse here, reports results from emission inventory of 2000 produced by the Agenzia Regionale per la Protezione Ambientale (ARPA), Tuscany. The basic model for an emission estimate is the product of (at least) two variables, for instance:

- an activity statistic and a typical average emission factor for the activity, or
- an emission measurement over a period of time and the number of such periods emissions occurred in the required estimation period.

For example, to estimate annual emissions of sulphur dioxide in grams per year from an oil-fired power plant you might use, either:

- annual fuel consumption (in tonnes fuel/year) and an emission factor (in grams SO_2 emitted/tonne fuel consumed), or
- measured SO_2 emissions (in grams per hour) and number of operating hours per year.

In practice, the calculations tend to be more complicated but the principles remain the same.

Emission estimates are collected together into inventories or databases which usually also contain supporting data on, for example, the locations of the emissions sources; emission measurements where

DOI: 10.1002/env

available; emission factors; capacity, production or activity rates in the various source sectors; operating conditions; methods of measurement or estimation; and so on.

Emission inventories may contain data on three types of source, namely point, area and line. However, in some inventories all data may be on area basis—region, country, subregion and so on.

Point sources: Emission estimates are provided on an individual plant or emission outlet usually large and in conjunction with data on location, capacity or throughput, operating conditions and so on. The tendency is for more sources to be provided as point sources, as legislative requirements extend to more source types and pollutants as well as more openness provides more such relevant data.

Area sources: Smaller or more diffuse sources of pollution are provided on an area basis either for administrative areas, such as counties, regions and so on, or for regular grids (for instance the EMEP 50×50 km grid).

Line sources: In some inventories, vehicle emissions from road transport, railways, inland navigation, shipping or aviation and so on are provided for sections along the line of a road, railway-track, sea-lane and so on.

In our study all estimates are referred to municipalities, then they can be seen as area sources type. We use the geographical centroid of the municipality territory as geographical reference. The total emission we are going to use is the summation of all available sources emission estimates in each municipality.

Pressure indicators are estimates referred to the year 2000 of potential loadings of the most relevant pollutants in each municipality. Estimates are available by emission source[†] according to the CORINAIR (CO-oRdinated INformation on the Environment in the European Community, AIR), our choice is to work with 2000 total estimated loadings of: Nitrogen Oxides (mg), Sulphur Oxides, Volatile Organic Compounds (VOC) (mg), Particulate Matter (mg), Carbon Monoxide (mg), Carbon Dioxide (mg) and Benzene (kg).

3.3. *Waste: urban and dangerous*

The Regulation 2150/2002/EC on Waste Statistics requires Member States to release data on the generation of waste for each economic activity from A to Q of the Statistical Nomenclature of Economic Activities in the European Community (NACE, Rev. 1) and for households. Also data related to waste management activities (recovery and disposal operations) are requested.

While data related to waste generation have to be released at national level, data on recovery and disposal operation are requested at NUTS1 (Nomenclature of Territorial Units for Statistics—region aggregations). Member States will furnish data every second year after the first reference year (2004), and within 18 months from the end of the reference year.

In Italy waste data are obtained yearly through a Compulsory Declaration (MUD)[‡] introduced by the Italian law no. 70/1994. MUD represents an administrative source and produces a great number of statistical information concerning collection, treatment and disposal related to municipal waste and waste produced by economic activities. This declaration requests, every year, all Italian municipalities and local units producing and/or managing waste to fill in the above-mentioned questionnaire. This information allows Italy to meet most regulation requirements. In detail available indicators are:

[†] Public Power Stations, Cogeneratio, Combustion from Tertiary and Agriculture, Industrial Combustion, Productive processes, Extraction and distribution of fossil combustible, Use of solvent, Road transportation, Other mobile sources, Waste treatment, Changes in soil composition due to Forestry and Agriculture, Nature.

[‡] Modello Unico di Dichiarazione ambientale.

DOI: 10.1002/env

- urban wastes (ton/year);
- separate collection (ton/year);
- dangerous waste (ton/year).

We consider these three indicators in their per capita version as pressure indicators. The reference year is 2001 and the spatial unit is the municipality. All pressure indicators are aligned in space.

4. EDA OF ENVIRONMENTAL MATRICES

As illustrated in the previous section, available data are of very diverse nature. As far as pressure indicators are concerned, data are aligned in space and misaligned in time. Most indicators are obtained through model-based estimation procedures, they have the most diverse range and measurement scales. In order to minimise problems raised by these heterogeneous data, we adopt the following strategy: first we explore each indicator behaviour, producing maps and exploratory analysis aiming to a better understanding of each framework; then we use a rank-based approach to produce a first classification inside each (covariate) subspace. The choice of rank is justified by the robustness of these classes of techniques with respect to 'low' quality data, diverse measurement scale and heterogeneity. In other words we assume that relative relationships between municipalities are constant in time and 'well' represented by our data.

As multiple indicators define each environmental dimension, we have to choose among several synthesis techniques. If we first rank municipalities according to each indicator, we can choose the average rank (or a weighted average) as final output. However, other choices are possible. In this work we investigated the use of the Wroclaw taxonomic technique and some of its variations. Roughly speaking the taxonomic technique is based on the definition of an *ideal unit* using, for instance, all indicators maxima or pollutants limits established by law (all properly standardised), which becomes the reference point from which we compute the Euclidean distance of each other unit. Then units are ranked according to their distance from this ideal unit. In this paper the ideal unit is built through indicators maxima, as a consequence a municipality with a low rank is characterised by values of indicators close to their maximum.

Variations of this technique build the ideal unit through principal component techniques (PCA). This choice is justified by the attempt to eliminate the influence of variables correlation (being PCA factors orthogonal). In other words by projecting all units on the principal plans and choosing the ideal unit in this reference system, we can compute distances of units removing variables correlation.

4.1. Air

Indicators belonging to the air matrix show a very wide range of variation. We adopt a five-class values clustering in order to map municipalities (Figure 2), classes extremes are based on pollutants percentiles (0, 0.25, 0.5, 0.75, 0.95, 1.00). In terms of linear correlation some strong linkages appears as it can be seen in Table 1, Benzene, CO and Cov; CO_2, NO_x, SO_x and Psf show high correlation. This is due to the emission estimation process, as these groups of pollutants are produced by the same sources.

We build municipalities ranking for each indicator and we aim to produce one ranking from the seven we obtain. As already mentioned we compare several synthesis techniques: the average rank, the taxonomic method and a variation of the taxonomic method based on a principal components-type

DOI: 10.1002/env

Table 1. Air matrix: linear correlation between pollutants

	Benzene	CO	CO_2	NO_x	SO_x	Psf	Cov
Benzene	1.00						
CO	0.96	1.00					
CO_2	0.57	0.73	1.00				
NO_x	0.75	0.87	0.92	1.00			
SO_x	0.42	0.59	0.93	0.87	1.00		
Psf	0.74	0.86	0.89	0.94	0.85	1.00	
Cov	0.90	0.86	0.41	0.65	0.28	0.64	1.00

technique accounting for spatial variation (Jona Lasinio, 2001). In order to choose the final ranking we compare results by computing the *W* Kendall's index. All techniques return ranking with high concordance degree ($W \cong 0.98$). Our choice is to use the taxonomic ranking as it produces the most coherent ordering according to our previous knowledge.

4.2. *Waste*

Due to the type of information collection, the waste matrix is not severely affected by missing data, only a very small percentage (11 out of 287 municipalities) of municipalities did not communicate their waste production. We adopt a stratified imputation approach in which municipalities are clustered according to their size (resident population), then the mean of each strata is used as a imputed value. This choice is justified by the strong linear relation existing between urban waste amount and municipality size.

As far as dangerous waste is concerned, the missing data (only two municipalities are missing) imputation has been based on the same technique with reference population the number of employees by sectors producing dangerous waste.[§]

As a first exploratory approach we classify Tuscan municipalities according to quantiles of each considered indicator. From Figure 3 we can see that most municipalities with a large per capita waste production adopt a strong policy of separate collection. On the other hand the per capita dangerous waste production shows a different behaviour, with a large number of municipalities producing only a very small amount. In terms of spatial autocorrelation only the Urban waste indicator results in a significant Moran test performed adopting the same neighbourhood structure chosen for the Air matrix, while Separate collection shows a significant spatial autocorrelation only if a randomisation technique is used to estimate the variance.

Again we build municipalities ranking for each indicator and we aim to produce one ranking from the three we obtain. Here we have to consider the reduction to the pressure given by a 'high' level of separate collection. We compare several synthesis techniques: the simple average rank, a weighted average rank where we assign larger weights to larger separate collection and the taxonomic method. In order to choose the final ranking we compare results by computing the *W* Kendall's index. All techniques return ranking with high concordance degree ($W \cong 0.99$). We are going to use the average

[§] Manufacturing, construction and demolition, energy production, agriculture, mining and quarrying.

DOI: 10.1002/env

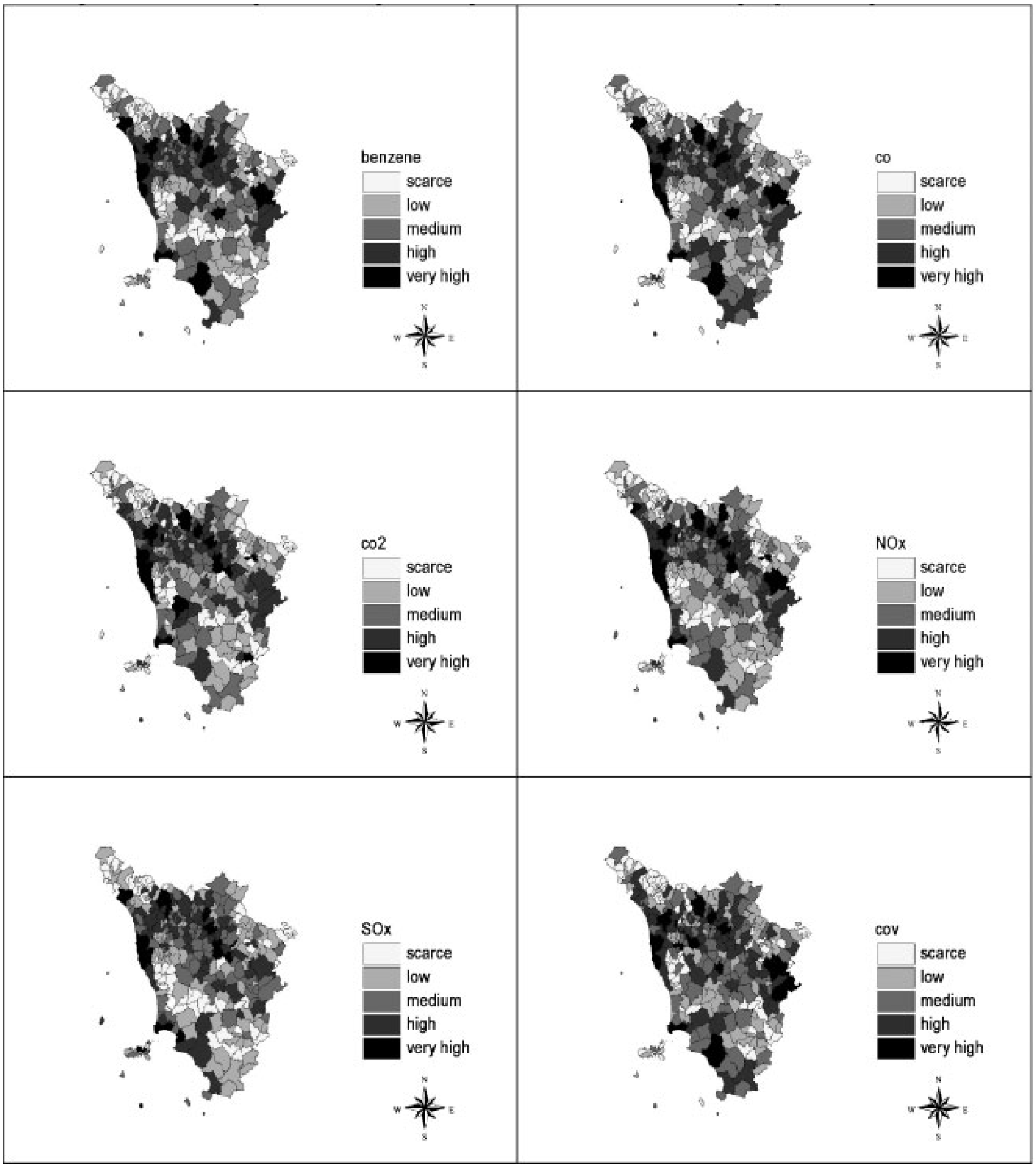

Figure 2. Air matrix: pollutants maps. Municipalities are clustered according to pollutants percentiles. *Benzene* [144.79–1080.855) scarce, [1080.855–2328.14) low, [2328.14–4963.5) medium, [4963.5–18716.752) high, [18716.752–150049) very high; *CO* [48–269.495) scarce, [269.495–558.56) low, [558.56–1083.06) medium, [1083.06–3976.339) high, [3976.339–27079) very high; *CO_2* [1146.5–9366.665) scarce, [9366.665–25750.82) low, [25750.82–64522.065) medium, [64522.065–371846.762) high, [371846.762–7984950) very high; *NO_x* [5.53–32.43) scarce, [32.43–102.43) low, [102.43–281) medium, [281–1055.6) high, [1055.6–12938) very high; *SO_x* [0.2–1.58) scarce, [1.58–5) low, [5–15.95) medium, [15.95–122.177) high, [122.177–39111) very high; *Cov* [15.01–120.07) scarce, [120.07–219.07) low, [219.07–440.675) medium, [440.675–1960.472) high, [1960.472–12575) very high; *Psf* [3.25–14.91) scarce, [14.91–26.83) low, [26.83–54.84) medium, [54.84–161.091) high, [161.091–1261) very high

DOI: 10.1002/env

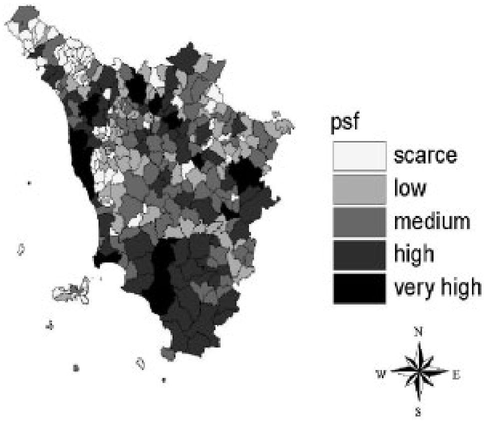

Figure 2. *Continued.*

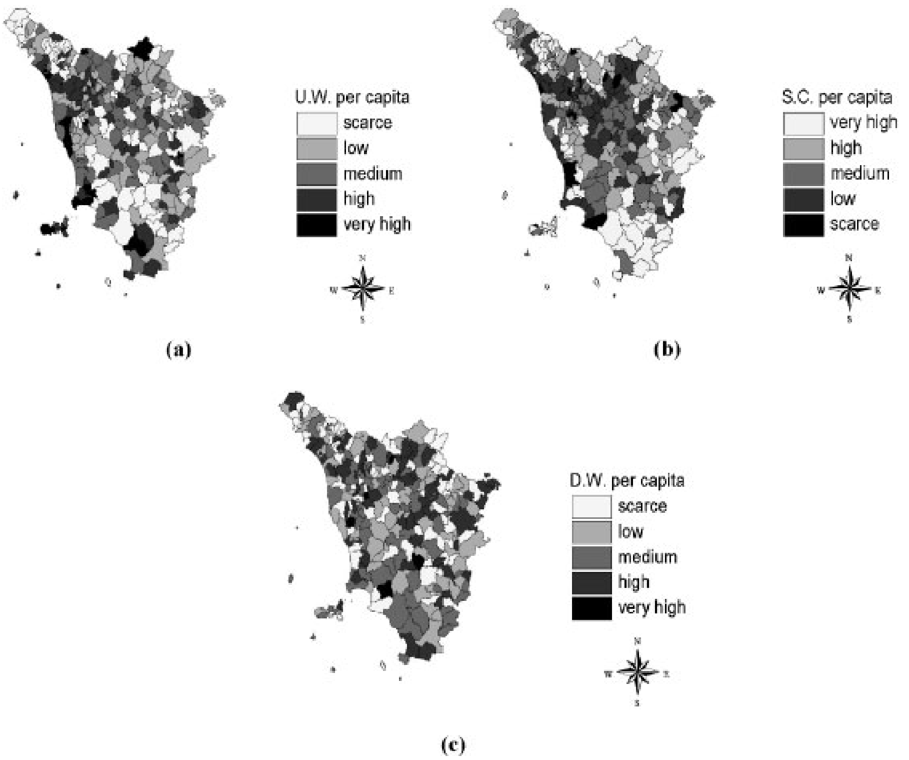

Figure 3. Quantiles classification of (a) per capita Urban waste [Per capita Urban Waste quantiles corresponds to [0.3–0.489) scarce, [0.489–0.558) low, [0.558–0.639) medium, [0.639–0.935) high, [0.935–1.67) very high]. (b) Per capita Separate collection [Ratio Separate collection and Urban Waste [0.002–0.081) scarce, [0.081–0.124) low, [0.124–0.159) medium, [0.159–0.26) high, [0.26–45.287) very high]. (c) per capita Dangerous Waste [Per capita Dangerous Waste 0 none, [0–0.035) scarce, [0.035–0.076) low, [0.076–0.19) medium, [0.19–5) high, [5.1–41.489) very high]

ranking as in this case this synthesis technique is the one producing the most interpretable municipalities ordination.

4.3. Water

As far as the pressure indicators are concerned, no missing data are present. For this set of indicators we proceed to describe their spatial behaviour by mapping municipalities classified, as in the other matrices, according to percentiles of each variable. In Figure 4 results show that nitrogen and phosphate pressure is very similar for most municipalities, a simple correlation analysis returns a 0.97 linear correlation between nitrogen and phosphate, 0.69 between nitrogen and organic and only 0.39 between organic and phosphate. Spatially, they all show a clustering behaviour pointing to the southern part of the region as the one potentially under higher pressure. We carried out Moran tests (two sided) to check for spatial autocorrelation. Nitrogen and phosphate show a high level of spatial autocorrelation, while

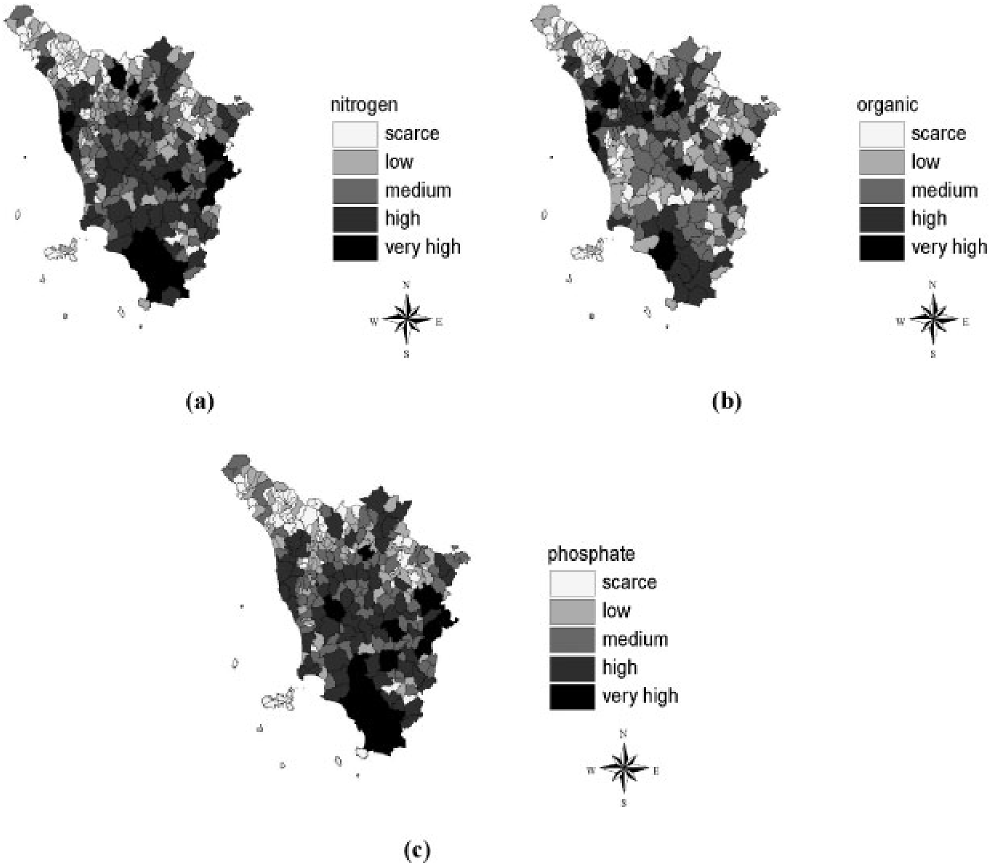

Figure 4. Water estimated potential loadings (a) Nitrogen [Nitrogen loadings [6.3–89.55) scarce, [89.55–183.50) low, [183.50–331.60) medium, [331.6–740.75] high, [740.75–2405.5) very high]. (b) Organic [Organic loadings [827.49–10037.83) scarce, [10037.83–22840.69) low, [22840.69–46880.88) medium, [46880.88–145493.27] high, [145493.27–786510) very high]. (c) Phosphate [Phosphate loadings [0.8–31.8) scarce, [31.8–70.9) low, [70.9–147.8) medium, [147.8–339.91] high, [339.91–1066.6) very high]

DOI: 10.1002/env

tests for organic pressure do not return allow to reject the null hypothesis of zero spatial autocorrelation. The neighbourhood structure adopted in tests is the same applied for both Air and Waste matrices.

As for the other two matrices we produce a final classification based on ranks. We follow the same scheme illustrated above. Again, in this case too, we obtain highly concordant ranking. Here as in the Air matrix case, we choose the taxonomic ranking as it produces the most coherent municipalities ordination.

5. CLASSIFICATION RESULTS

After obtaining the final ranking for each environmental matrix, we feed them to our *gate*, built by using the mixture model described in Section 2. In detail, response variables y_{id} in Equation (1) are the ranks of municipality i ($i = 1, \ldots, n$) in matrix d ($d = 1$, 2, 3, air, water and waste).

We implemented a Gibbs-Metropolis sampler algorithm[¶] in order to simulate the posterior distribution given in Equation (2). In particular the MCMC algorithm uses Gibbs steps to update latent labels and weights, while Metropolis steps, with, respectively, Gaussian random walk proposals and log-Gaussian random walk proposals, to update mean and precision parameters. We iterate the algorithm for 15 000 sweeps with a burn-in period of 10 000 iterations. The algorithm performances have been estimated to about 30 min on a Pentium 4, 3 GHz CPU. We estimate model parameters and latent labels using the last 5000 iterations. Convergence has been verified through the Gelman-Rubin statistics (Robert and Casella, 1999).

In order to choose the optimal number of classes we run our model with respect to several values k ($k = 2, \ldots, 8$) and we compute the Bayesian Information Criterion (BIC) for each run. This choice is justified by the complexity of both model and data (McLachlan and Peel, 2000) and this criterion is appropriate when data complexity must be taken into account.

In Table 2 we report BIC and log-likelihood values obtained from these runs. The 'optimal' classification, according to BIC, is given by $k = 4$.

In Table 3 class centroids are reported with their MCMC standard deviations. Classes are easily interpretable, remember that smaller the rank higher the value of pressure indicators. Then Class 1, as it includes all municipalities with a low rank in each region of the covariate space, can be labelled 'high pressure'; air, water and waste all require attention. Class 2 can receive a 'medium–high pressure'

Table 2. BIC and log-likelihood statistics for the choice of classes number

Classes	Log-likelihood	BIC
2	−211.460	502.160
3	−166.494	451.848
4	−139.287	437.054
5	−121.647	441.394
6	−108.289	454.298
7	−100.690	478.720
8	−92.710	502.380

[¶] The algorithm has been programmed in Fortran.

DOI: 10.1002/env

Table 3. Classes centroids, in parenthesis the MCMC standard deviation

Classes	Air	Water	Waste
1	94.17 (1.59)	99.42 (1.31)	74.17 (1.52)
2	79.30 (2.64)	78.56 (2.41)	211.33 (1.49)
3	225.24 (1.25)	228.12 (1.37)	72.02 (1.37)
4	206.02 (3.38)	197.62 (3.68)	220.98 (1.95)

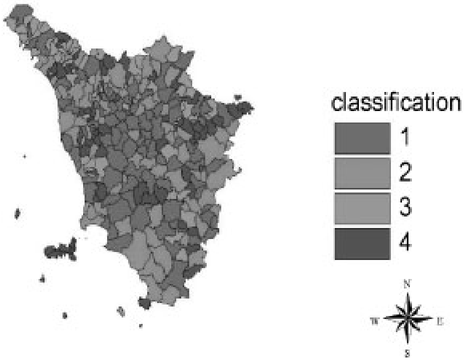

Figure 5. Tuscany municipalities: final classification map

label; air and water both require more attention than in Class 1, but waste is a nonproblematic sector. Class 3 is a ‘medium–low pressure’ class, only the waste treatment and production require attention. In Class 4 (‘low pressure’), all indicators point to a relatively acceptable situation.

In Figure 5 we map classification results. The majority of municipalities belong to Classes 1 and 2, recall that municipalities in Class 2 need more attention than the ones in Class 1 with respect to air and water pressure. Low pressure municipalities are spread all over the region; it is interesting to notice that the islands and the naturalistic protected area around Orbetello (south part on the coast) can be coherently classified as subject to a low environmental pressure.

6. CONCLUSION AND FURTHER DEVELOPMENTS

In Section 1 we describe our procedure as a combined classifier in which the gate is made by a mixture-based classifier and ‘researcher intervention’. This last aspect is given by the decision we made: which classification was ‘best’ for each environmental matrix.

The choice of ranks and techniques to synthesise them is the output of ‘human action’ in this general protocol. Clearly several other choices are possible and require investigation. However, with this first attempt, we obtained a coherent municipality classification using a flexible approach that solves dimensionality and complexity problems typical of ERA procedures. The semi-automatic nature of our approach overcomes most ‘conceptual’ questions raised by the definition of environmental risk.

We adopted Bayesian mixture classifiers, but there are other classification techniques stemming from Pattern Recognition literature that could be used in building the gate of our meta-classification, such as Neural Network, Support Vector Machines and so on (see Center, 2001 for a comparison).

However, all the above-mentioned procedures require the choice of a training sample, choice to which they are quite sensitive. In addition, mixture classifiers allow the researchers to give an objective interpretation of the output classes. In fact, we can objectively classify spatial units and then, through the parameters estimates of the mixture components, describe the class profiles and establish which environmental sector requires intervention or attention. Furthermore in each step of our procedure data quality, data heterogeneity and uncertainty evaluation are taken into account. Notice that our interpretation of classes is based on the knowledge that municipalities with low rank do not fulfil environmental regulation standards.

In order to choose the optimal classification we ran our model with respect to several dimensions (the number of classes k) and then we computed the BIC for each run in order to evaluate the fitting and complexity of the model. A possible alternative to our approach is represented by the reversible jump Markov chain Monte Carlo (RJMCMC) technique (see e.g. Green, 1995; Richardson and Green, 1997; Viallefont *et al.*, 2002) and its 'birth-death process' variant (see Stephens, 2000), which allows the model to choose also the possible number of classes k directly through the MCMC simulations. That approach is more general and flexible, but on the other hand, it introduces a larger complexity in the computation of the Bayesian mixture because the dimensionality of the problem now is considered as a new random quantity to be estimated. As a first attempt in the implementation of the classification model, we choose to favour a simple and operative data analysis than to deal with the further complexity that a RJMCMC approach surely introduces in the problem. However, a RJMCMC-based classification can represent a substantial part of our future research.

Another important issue in Bayesian modelling is the robustness of the results with respect to the tuning of the hyperparameters. Richardson and Green (1997) observed that a Bayesian mixture model could produce a posterior distribution influenced by the prior choice of the hyperparameters. This fact is true overall in a RJMCMC framework; indeed this problem affects mainly the posterior distribution profiles of the number of the classes k, see also McLachlan and Peel (2000). In our application the effect is quite weak, due also to the fact that the input data of the mixture are robust transformations of the initial data, as the ranks are. Anyway an empirical tuning of the hyperparameters of the mixture over a small grid of values has not shown significant changes in the final classification.

Several other aspects need further investigation. For instance it may be of interest to build a weighted version of the mixture classifier, in which each dimension (environmental matrix) receives a different weight according to some relevance criterion modulated by a probabilistic model (in order to account for criterion uncertainty).

A more practical development will be to apply our approach to a larger number of dimensions, possibly including human health-related variables and soil quality indicators. As far as status indicators classification, the main problem to be solved is data availability. At present only waste and water status indicators are available.

As a final remark, it is of interest to investigate the generalisation of our approach by defining a unique model able to enclose both classification steps. The main problem here is the complexity of the phenomena combined with the important point of data quality. In order to implement any statistically complex structure, such as a Bayesian hierarchical model or a probabilistic clustering procedure or any other pattern recognition procedure on each environmental dimension, the researchers need good data. To our knowledge, a very robust statistical approach with respect to data quality is the rank transformation. Then, one attempt could be to apply a stochastic mechanism to generate the partial ranks of municipalities inside each environmental dimension.

In general terms, we can formalise this idea as follows. Let $\mathbf{X}_{id}$ be the vector of pressure indicators for the i-th municipality in the d-th environmental dimension. Then we can assume that

DOI: 10.1002/env

$X_{id} \sim P(X_{id}|\theta_{id})$ with respect to a proper probability model P, where θ_{id} is a latent (scalar) variable denoting the *real state of nature* of the unit i in the dimension d. The final ranking of units is performed on the estimated values of θ_{id} in a *hidden model* framework. The main challenge here is the definition of a coherent probability model P. Indeed we have to account for data time misalignment, indicators heterogeneity, missing values imputation and so on. This is going to be a point to be developed in our future research.

ACKNOWLEDGEMENTS

We would like to thank Michela Leone, Alessia Marzano and Elisa Meko whose Masters Thesis work greatly helped in the development of this paper. The exploratory data analysis and the ***R*** implementation of a large part of this research have been derived form their work. A special thanks goes to the ARPA (Agenzia Regionale per la Protezione Ambientale) Tuscany staff who kindly gave us the data. In particular we thank Danila Scala for the many suggestions and explanations on the data. This work has been supported by the MURST COFIN project 2002–2004 'Statistics for environmental management: applications to space-time data series and environment-health interactions'.

REFERENCES

Brutti P, Fabi F, Jona Lasinio G. 2002. Speaker recognition: a proposal for a meta-analysis based on hierarchical combination of classifiers. *Statistica* **LXI**(3): 455–473.

Center J. 2001. Relating Bayesian mixture-model classifiers to other popular pattern classifiers. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 20th International Workshop*. American Institute of Physics. Gif-sur-Yvette, France.

Fairman R, Mead CD, Williams WP. 1999. Environmental Risk Assessment—Approaches, Experiences and Information Sources Environmental, Issue Report no. 4. EEA, European Environment Agency. http://reports.eea.eu.int/GH-07-97-595-EN-C2/en/riskindex.html

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.

Green PJ, Richardson S. 2002. Hidden Markov models and disease mapping. *Journal of the American Statistical Association* **97**(460): 1055–1070.

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. 1991. Adaptive mixtures of local experts. *Neural Computation* **3**(1): 79–87.

Jona Lasinio G. 2001. Modeling and exploring multivariate spatial variation: a test procedure for isotropy of multivariate spatial data. *Journal of Multivariate Analysis* **77**: 295–317.

Jona Lasinio G, Rossi C, Bove T. 2004. The speaker recognition problem. In Proceedings of the XLII Scientific Meeting of the Italian Statistical Society, Bari, Italy.

Jordan MI, Jacobs RA. 1994. Hierarchical mixtures of experts and EM algorithm. *Neural Computation* **6**(2): 181–214.

McLachlan G, Peel D. 2000. *Finite Mixture Models*. John Wiley & Sons, Inc.: New York.

Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society* Ser. B, **59**(4): 731–792.

Robert CP, Casella G. 1999. *Monte Carlo Statistical Methods*. Springer: New York.

Royal Society. 1992. *Risk Analysis, Perception and Management*. The Royal Society: London.

Stephens M. 2000. Bayesian analysis of mixtures with an unknown number of components—an alternative to reversible jump methods. *Annals of Statistics* **28**(1): 40–74.

Suter GW II. 1993. *Ecological Risk Assessment*. Lewis Publishers: Boca Raton.

Van Leeuwen CJ, Hermens JLM (eds). 1995. *Risk Assessment of Chemicals: An Introduction*. Kluwer Academic Publishers: Dordrecht.

Various Authors. 1991. *Valutazione dei carichi inquinanti potenziali per i principali bacini idrografici italiani*. Quaderno no. 90, CNR-IRSA. Consiglio Nazionale delle Ricerche, Istituto di Ricerca sulle Acque (in Italian).

Various Authors. 2003. *Artemisia 2. Uno strumento per valutare gli effetti ambientali e sanitari degli inquinanti aeriformi emessi da insediamenti produttivi e per indirizzare la scelta di nuovi siti, anche ai fini dell'applicazione della direttiva 96/61 CE sulla prevenzione e riduzione integrata dell'inquinamento (IPPC). Applicazione all'area di Milazzo*. Research Report. ENEA, Ente per le Nuove tecnologie, l'Energia e l'Ambiente (in Italian).

Viallefont V, Richardson S, Green PJ. 2002. Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics* **14**: 181–202.