PROCEEDINGS

GRASPA•2015

GRASPA-SIS Biennial Conference

GRASPA–SIS: The Research Group for Environmental Statistics of the Italian Statistical Society

TIES European Regional Meeting

TIES: The International Environmetrics Society

Edited by: Alessandro Fassò and Alessio Pollice

Bari • 15-16 June 2015

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

GRASPA 2015

Introduction

A. Fassò[1], A. Pollice[2]

[1] *Dipartimento di Ingegneria gestionale, dell'informazione e della produzione, Università degli Studi di Bergamo, Viale Marconi 5, 24044 Dalmine (BG), ITALY; alessandro.fasso@unibg.it*
[2] *Dipartimento di Scienze Economiche e Metodi Matematici, Università degli Studi di Bari Aldo Moro, Largo Abbazia Santa Scolastica 53, 70124 Bari, ITALY; alessio.pollice@uniba.it*

GRASPA 2015 is the biennial conference of the Italian Research Group for Environmental Statistics (GRASPA-SIS). GRASPA is active since 1995 and has become a permanent working group of the Italian Statistical Society (SIS) since May 2013. GRASPA-SIS promotes statistical and interdisciplinary research in the field of environmental quality, safety and sustainability including air and water quality, epidemiology, climate, earth science and ecology. GRASPA 2015 is also the 2015 European regional conference of The International Environmetrics Society (TIES), it is sponsored by the Young section of the Italian Statistical Society and is a connected event of the Spatial Statistics 2015 Conference.

The meeting will have Tata Subba Rao (University of Manchester) and Adrian Bowman (University of Glasgow) as keynote speakers, ten invited tracks on various statistical and environmental topics and more than forty contributed papers. An extensive poster session with nominations for best poster awards will be held. A post-conference one-day short course on introducing flexible regression for environmental data will be given by Adrian Bowman (University of Glasgow).

A Book of Abstracts including 83 abstracts of keynote, invited and contributing authors will be printed, while these Conference Proceedings contain 31 invited and contributed short papers listed alphabetically according to the first author family name. Extended versions of a selection of invited and contributed papers will be considered for publication in special issues of the Journal of Statistical Computation and Simulation (guest editors: A. Pollice, G. Jona Lasinio) and Stochastic Evironmental Resarch and Risk Assessment (guest editors: E. Romano, J. Mateu, M.D. Ruiz-Medina).

Contents

F. Fedele, A. Pollice, A. Guarnieri Calò Carduccci, R. Bellotti
Spatial bias analysis for the Weather Research and Forecasting model (WRF) over the Apulia region

A. Ferruzza, D. Vignani, G. Tagliacozzo, S. Tersigni, A. Tudini
International frameworks for environmental statistics and their application to climate change related statistics

F. Finazzi, A. Fassò
Real-time detection of earthquakes through a smartphone-based sensor network

M. Franco-Villoria, R. Ignaccolo, A. Fassò, F. Madonna, B.B. Demoz
Collocation uncertainty in climate monitoring

C. Ghiglione, C. Carota, C.R. Nava, I. Soldani, S. Schiaparelli
Rarefaction and extrapolation with Hill numbers: a study of diversity in the Ross Sea

P. Giungato, P. Barbieri, F. Lasigna, G. Ventrella, S.C. Briguglio, A. Demarinis Liotile, E. Tamborra, G. de Gennaro
Integration of different electronic nose technologies in recognition of odor sources in a solid waste composting plant

K. Krivoruchko, D. Pavlushko
Improving R and ArcGIS integration

F. Manca, E. Loiacono, G.L. Cascella, D. Cascella
Energy-efficiency optimization of the biomass pelleting process by using statistical indicators

O. Nicolis
Environmental SmartCities: statistical mapping of environmental risk for natural and anthropic disasters in Chile

R. Pappadà, E. Perrone, F. Durante, G. Salvadori
A semi–parametric approach in the estimation of the structural risk in environmental applications

A. Rarugal, R.M. Roxas-Villanueva, G. Tapang
Impact of climatic factors on acute bloody diarrhea, dengue and influenza-like illness incidences in the Philippines

E. Recchini
Official statistics for decision making: an environmental accounting case study related to biodiversity

E. Venezia
Environmental sustainable management of urban networks with the use of ICT: URBANETS project. The case of Gallipoli

P. Zanini
Multi-resolution and spatial Independent Component Analysis approaches for geo-referred and time-varying mobile phone data

Spatio-temporal modelling of zero-truncated disease patterns

O. Adegboye[1,*], D. Leung[2] and Y-G. Wang[3]

[1]*Department of Mathematics, Statistics & Physics, College of Arts & Sciences, Qatar University; o.adegboye@qu.edu.qa,*
[2]*School of Economics, Singapore Management University, Singapore; denisleung@smu.edu.sg*
[3]*School of Mathematics and Physics, University of Queensland, Queensland, Australia; you-gan.wang@uq.edu.au*
[*]*Corresponding author*

Abstract. *This paper focuses on the spatio-temporal pattern of Leishmaniasis incidence in Afghanistan. We hold the view that correlations that arise from spatial and temporal sources are inherently distinct. Our method decouples these two sources of correlations, there are at least two advantages in taking this approach. First, it circumvents the need to inverting a large correlation matrix, which is a commonly encountered problem in spatio-temporal analyses (e.g., Yasui and Lele, 1997) [3]. Second, it simplifies the modelling of complex relationships such as anisotropy, which would have been extremely difficult or impossible if spatio-temporal correlations were simultaneously considered. The model was built on a foundation of the generalized estimating equations (Liang and Zeger, 1986) [1]. We illustrate the method using data from Afghanistan between 2003-2009. Since the data covers a period that overlaps with the US invasion of Afghanistan, the zero counts may be the result of no disease incidence or lapse of data collection. To resolve this issue, we use a model truncated at zero.*

Keywords. *Generalized estimating equations; Overdispersion; Poisson; Spatio-temporal*

1 Introduction

Leishmaniasis is the third most common vector-borne disease and a very important protozoan infection. The disease is contracted through bites from sand flies, which are themselves not poisonous, but the parasitic Leishmania in its saliva can result in chronic and non-healing sores. Some of the risk factors identified include household construction materials, design, density and presence of the disease in the neighborhoods and high rodent infestations. The impact of environmental influences on Leishmaniasis cannot be ruled out and human activities play a significant role in the dispersion of the vectors thereby changing the geographical distribution of the disease.

The present study was motivated by Leishmaniasis cases in the provinces of Afghanistan between 2003 and 2009. One of the most challenging issues in modelling spatio-temporal data is the choice of a valid and yet flexible correlation (covariance) structure. The correlation structures fall into one of two types: separable in which case it is assumed that the space-time correlation can be written as a product of a correlation for the space dimension and one for the time dimension or non-separable where the space-time correlation is modelled as a single entity. Mostly, space-time correlations are considered jointly, a

step that we believe is unnecessary or unrealistic in our data.

In this study we shall decouples these two sources of correlations, an approach that separates the modelling of the space- and time-correlations. There are at least two advantages in taking this approach. First, it circumvents the need to inverting a large correlation matrix, which is a commonly encountered problem in spatio-temporal analyses. Second, it simplifies the modelling of complex relationships such as anisotropy, which would have been extremely difficult or impossible if spatio-temporal correlations were simultaneously considered.

Our method is based on the framework of generalized estimating equations (GEE) where the spatial dependency is accounted for by re-weighting the standard GEE so that locations that are highly correlated with each other would receive less weight. Apart from the spatial dependency in our data, the data is also characterized by a high percentage of zero disease counts which introduced over-dispersion. Since the data covers a period that overlaps with the US invasion of Afghanistan, the zero counts may be the result of no disease incidence or lapse of data collection. It is often practiced to truncate the values that are bigger than a constant to overcome over-dispersion [2]. The analysis of truncated often arises from a subsidiary set of results that treat a practical problem of how data are gathered and analyzed and incompleteness of this data requires special estimators of the regression coefficients. To resolve this issue, we use a model truncated at zero.

The rest of the paper is structured as follows. Section 2, describes the materials and methods that will be used in the study. In Section 3 we shall give the results of the data analysis and conclude the paper

2 Data Sources and Methods

2.1 Data Sources

The data used in this study were monthly cases of Leishmaniasis reported to the Afghanistan Health Management Information System (HMIS) under the National Malaria and Leishmaniasis Control Programme (NMLCP) of the Ministry of Public Health (MoPH). The data consists of 148,945 new cases of Leishmaniasis from 20 provinces in Afghanistan between 2003 and 2009 (of these, 41,072 occurred in 2009). We used satellite-derived environmental data- Normalized difference vegetation index (NDVI), land surface temperature (LST) and rainfall as explanatory variables.

2.2 Model Formulation and Parameter Estimation

We begin by considering the disease counts $\mathbf{y} = (y'_{1.},...,y'_{S.})^T$ and observed covariates at different locations $\mathbf{X} = (x'_{1.},...,x'_{S.})^T$ as a set of longitudinal data over S spatial locations. Let $\mathbf{y}$ be independent and assumed to follow a Poisson model and stacked as a $S \times T$ vector. The covariance matrix of y is $\tilde{V}$ and $\tilde{v}^{-1}_{st,s't'}$ is the (st,st')-th element of $\tilde{V}^{-1}$, the dimension of $\tilde{V}$ is $ST \times ST$.

For the dataset we are working with, $S = 20$ represents the number of provinces and $T = 7$ represents the number of years with recorded data. Using the monthly data, then $T = 84$ and so $S \times T = 20 \times 84 = 1680$ and therefore $\tilde{V}$ would be a matrix that cannot feasibly be handled. Moreover, the correlation between $y(s,t)$ and $y(s',t')$ often does not have any practical meaning. For a fixed s, $v_{s,tt'}$, $t,t' = t_1,...,t_T$ are the elements of the variance covariance matrix of disease counts between times.

The modelling is a 2-step process, we first needed to find the variance covariance matrix, $v_{s,tt'}$ and spatial weight, $\tilde{w}_{ss'}$. We compute empirical temporal variograms at different spatial locations and then average all temporal variograms with the same temporal lag. We applied the empirical semivariogram based on the Pearson residuals and fitted a parametric semivariogram models. For two different times, say t,t', that are $t = |t - t'|$ months apart, the correlation between the two times, t,t' could be written as:

$$C(t,t') = C^0_T(t) \tag{1}$$

where $C_T^0(t) = e^{(-\theta t)}$ is the temporal covariance function with months apart. The parameters $\theta = \tau^2, \sigma^2, \phi$ represents the nugget, sill, and range, respectively.

In order to model spatial correlation and overdispersion, we assume there is a nonnegative weakly stationary latent process e and conditioned on this process, the y's are independent and follow a log-linear model given below. Consider the following; suppose we remove all $y_{st} = 0$, then conditioned on $y_{st} > 0$, we have $E(y_{st}|e_{st}) = c\mu_{st}(\beta)e_{st}$, and $var(y_{st}|e_{st}) = [c\mu_{st}(\beta) + c(1-c)\mu_{st}(\beta)^2]e_{st}$ where $c = 1/[1-\exp(-\mu_{st}(\beta))]$, leading to

$$E(y_{st}) = c\mu_{st}(\beta) \equiv \phi_{st}(\beta), \tag{2}$$

$$var(y_{st}) = c\mu_{st}(\beta) + c(1-c)\mu_{st}(\beta)^2 + c^2\mu_{st}(\beta)^2\sigma^2. \tag{3}$$

where β are unknown parameters. We assume $\mathrm{E}(e_{st})$ to be 1 so that $\mu_{st}(\beta)$ represents the marginal mean of y_{st}.

Let $\doteq\{d(s,t) = d_{st}\}_{S\times T}$ be a matrix of indicators such that $d_{st} = 1$ if $y_{st} > 0$ and $d_{st} = 0$ otherwise. Note that $y_{st} = 0$ could mean the count was zero or count was not taken. For a particular set of spatial weight $\tilde{w}_{ss'}$, the spatial GEE conditioned only on those observations with $y_{st} > 0$ can be written as

$$\tilde{U}(\beta,\alpha) \equiv \sum_{s=s_1}^{s_S}\sum_{s'=s_1}^{s_S}\sum_{t=t_1}^{t_T}\sum_{t'=t_1}^{t_T}\frac{\partial\phi_{st}}{\partial\beta^T}d_{st'}\tilde{w}_{ss'}v_{s,tt'}^{-1}\{y_{st'} - \phi_{st'}\} = 0, \tag{4}$$

where $v_{s,tt'}$ is the t,t-th element of V_s, the covariance matrix of $y_{s\cdot}$. The matrix V_s can be expressed as $A_s^{1/2}R_s(\alpha)A_s^{1/2}$, where $A_s = \mathtt{diag}[c\mu_{s1}(\beta) + c(1-c)\mu_{s1}(\beta)^2 + c^2\mu_{s1}(\beta)^2\sigma^2, ..., c\mu_{sT}(\beta) + c(1-c)\mu_{sT}(\beta)^2 + c^2\mu_{sT}(\beta)^2\sigma^2]$ and $R_s(\alpha)$ is a matrix with its (t,t')-th element representing the correlation between times t and t' at location s.

Our primary interest lies in the parameters β but we also must deal with the nuisance parameters α. Let $R(\alpha)$ be a 84×84 matrix where α contains the parameters (θ) estimated via weighted least square method. The parameters are estimated via a Newton-Raphson iteration method. To solve for (α,β) jointly. Let $\hat{\beta}_k$ and $\hat{\alpha}_k$ be the estimates of β and α at the k-th iteration. We first fitted a GEE with an independence working correlation structure, we then solve the estimating equation for α, and we then iterate until convergence. This step gives the values $v_{s,tt'}$. Denoting $\sum_{s,s',t,t'} \equiv \sum_{s=s_1}^{s_S}\sum_{s'=s_1}^{s_S}\sum_{t=t_1}^{t_T}\sum_{t'=t_1}^{t_T}$, we estimate an initial estimate $\hat{\beta}_0$ using (4) by assuming an identity matrix for $R_s(\alpha)$, equivariance, *i.e.*, $v_{s,tt'}^{-1} = 1$ and, spatial weight.

Then at iteration k,

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \left[\sum_{s,s',t,t'}\frac{\partial\phi_{st}(\hat{\beta}_k)}{\partial\beta^\tau}d_{st'}\tilde{w}_{ss'}v_{s,tt'}^{-1}(\hat{\beta}_k)\frac{\partial\phi_{st}(\hat{\beta}_k)}{\partial\beta^\tau}\right]^{-1}\left[\sum_{s,s',t,t'}\frac{\partial\phi_{st}(\hat{\beta}_k)}{\partial\beta^\tau}d_{st'}\tilde{w}_{ss'}v_{s,tt'}^{-1}\{y_{st'} - \phi_{st'}(\hat{\beta}_k)\}\right]. \tag{5}$$

Since we are using an AR(1) correlation structure, we take the slope of the linear regression of $\log(\hat{r}_{st}^k\hat{r}_{st'}^k)$ on $\log(|t-t'|)$ as $\hat{\alpha}_k$. We then iterate between (4) and (5) until convergence.

The standard errors for the $\beta's$ were obtained using large-sample properties.

3 Illustration

We shall illustrate our method using the Leishmaniasis cases data reported to the Afghanistan Health Management Information System (HMIS) of the Ministry of Public Health (MoPH) between 2003 and 2009. We observe higher disease incidence around the Kabul area (North Eastern). Similar patterns were observed in 2003-2008 (maps not shown here but are available on request). The monthly profile of cases of Leishmaniasis revealed two peaks in the disease occurrence in Afghanistan between 2003 and 2009 – January to March and September to December – which coincide with the cold period while July is the

Table 1: Parameter estimates together with the standard errors from GEE with different correlation structures of Leishmaniasis incidence in Afghanistan

Risk factors	$GEE_{Spatial}$	$GEE_{Temporal}$	$GEE_{Spatio-temporal}$
Intercept	-0.52289 (0.07342)	-9.09746 (0.08526)	-9.09818 (0.02206)
Altitude (*m*)	-0.00012 (0.00023)	0.00026 (0.00001)	0.00026 (0.00001)
Temperature (°C)	-0.42460 (0.00022)	-0.00118 (0.00035)	-0.00113 (0.00017)
Precipitation (*Inches*)	1.58830 (0.00785)	-0.03920 (0.00066)	-0.03895 (0.00210)
Wind (*Knot*)	0.53639 (0.00528)	0.02089 (0.01566)	0.02078 (0.00112)
2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$	69.33	87.054	19.38
AIC	103.112	179.39	46.511

hottest month and March is the wettest month. The time series plot for the number of Leishmaniasis cases reveal upward trend and regularly repeating patterns of highs and lows related to the months of the year which suggests seasonality in the data. The variogram of space-time autocorrelation is obtained by considering time as discrete. This method models the cross-variograms between data with time replication (months/years) and captures the variability in space and time. We hold the view that correlations arising from spatial and temporal sources are inherently distinct. Our method makes it possible to combine the specific provincial rate with the influence of the spatial neighborhood. Three different models were fitted namely; spatial only, temporal only and spatio-temporal model. In Table 1, perhaps the most distinctive results are from the model with spatial correlation; the model parameter estimates are remarkably different from others. The result may not be surprising as it has been assumed that the correlation remains the same across time. This also suggests that spatial correlation only may not be sufficient for the data, because it involves the specification of spatial correlation across time. The results have shown that the specified spatio-temporal function is more suitable and appropriate for this data (smaller 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$). Moreover, the model with the spatio-temporal correlation function significantly improves the model fit when compared to other specifications, as judged by the smaller AIC. Although the parameter estimates from both temporal and spatio-temporal models are similar, significant differences can be observed in their precision estimation. The technique used is this study allow for correct specification of correlation structures to improve the efficiency of the GEE method. The Leishmaniasis data presented several problems with modelling issues, ranging from correlation/covaraince specification to issues with "imputed" or "non true" zeros. The high percentage of zero disease counts may be the result of no disease incidence or lapse of data collection. Moreover, the dependency in the data may be a result of spatial variation, temporal or both. To resolve this issue, a renowned method was used to address the many issues that the data presented in a very novel way. A model truncated at zero was fitted while allowing for dependency in the data via a working correlation matrix using the technique of GEE.

References

[1] Leung, D., Wang, Y.-G., and Zhu, M. (2009). Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics* **10**, 436–445.

[2] Saffari, S. E., Adnan, R., and Greene, W. (2011). Handling of over - dispersion of count data via truncation using poisson regression model. *Journal of Computer Science and Computational Mathematics* **1**.

[3] Yasui, Y. and Lele, S. (1997). A regression method for spatial disease rates: An estimating function approach. *Journal of the American Statistical Association* **92**, 21–32.

Looking for changepoints in spatio-temporal earthquake data

L. Altieri[1*], D. Cocchi[1], F. Greco[1], J. B. Illian[2] and E. M. Scott[3]

[1] *University of Bologna, Department of Statistics; {linda.altieri, daniela.cocchi, fedele.greco}@unibo.it*
[2] *University of St Andrews, CREEM, School of Mathematics and Statistics; janine@mcs.st-and.ac.uk*
[3] *University of Glasgow, School of Mathematics and Statistics; marian.scott@glasgow.ac.uk*
[*] *Corresponding author*

Abstract. *This work presents an application of a new method for changepoint detection on spatio-temporal point process data. We summarise the methodology, based on building a Bayesian hierarchical model for the data and priors on the number and positions of the changepoints, and introduce two approaches to taking decisions on the acceptance of potential changepoints. We present the dataset collecting Italian seismic events over 30 years and show results for multiple changepoint detection. Finally, concluding comments and suggestions for further work are provided.*

Keywords. *earthquake data; changepoint analysis; spatio-temporal point processes; log-Gaussian Cox processes*

1 Introduction

This work provides an application of new methodology for changepoint analysis on spatio-temporal point process data as proposed in [1]. The case study consists of all Italian seismic events exceeding a specific magnitude recorded in the last 30 years.
The collected data are provided by INGV (the National Institute of Geophysics and Vulcanology) and are free to download at `http://terremoti.ingv.it/it/`. They are published in real time and cover all seismic events from January 1985 onwards. For each event, the spatial coordinates, the hypocentre depth and the magnitude are reported. Data come from 390 monitoring stations located over the Italian territory, which operate 24 hours a day, 7 days a week. We analyse a set of 19774 events of magnitude 2.5 and above (earthquakes below this limit are not felt by people). The study period covers from January, 1985 to December, 2014. A map of the hypocentre locations is presented in Figure 1. We split the dataset into yearly patterns and obtain a time series of spatial point processes (where timepoints are years) with a number of seismic events ranging from 304 to 1592, with an average of 659 per year.
A changepoint analysis can answer many questions concerning the evolution of the seismic phenomenon over the Italian territory. Issues that need to be met are listed in many recent articles in the INGV website and highlight concerns about changes occurring in the distribution and magnitude of earthquakes. Since it is sensible to assume spatial correlation and temporal dependence among the events, the development of a methodology able to face such a complex dataset now allows these questions to be answered. Secondly,

as stated in [1], there is a need to provide a proper application of the new method. Indeed, the motivating case study presented in [1], though interesting, is limited as regards the length of the time series ($T = 15$) and the low number of events in some years. We aim here to show a more complex case study and to answer practical questions about changes in earthquake phenomena.

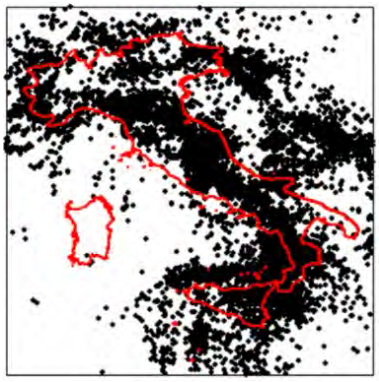


Figure 1: Seismic events of magnitude$\geq$2.5, 1985-2014.

2 Changepoint detection on spatio-temporal point processes

At every time point the datum consists of a realisation of a spatial point process, therefore different types of change over time may occur: a change in scale (expected number of points), in spatial distribution or in both. Moreover, in real situations issues of spatial dependence among points and temporal dependence within time segments must be considered. Recent work [1] has developed a new Bayesian method for the detection of an unknown number of temporal changes over a spatio-temporal inhomogeneous point process where spatial and temporal dependence within time segments are allowed. The validity of the method has been assessed in a thorough simulation study, and it has been shown to be able to detect different types of change. The use of INLA [4] to compute the segment marginal likelihoods makes the approach computationally tractable.
In a nutshell, the method consists in choosing a model and fitting it multiple times to the dataset assuming different changepoint positions. Every time a changepoint is assumed at a timepoint $\theta = 1, \ldots, T$, the data vector is split into two segments based on the changepoint location and the model is fitted separately to the two segments (independence across segments is assumed here). Two segment log-likelihoods values are obtained and summed to give the marginal log-likelihood conditional on θ. For different changepoint locations, a vector of log-likelihoods is computed. The posterior distribution of the changepoint location is obtained via the Bayes Rule by multiplying the log-likelihood vector for a vector of prior probabilities over the changepoint positions θ. Once a posterior probability is obtained for every time point, decisions must be made as to which changepoints are to be accepted. For a multiple changepoint search, we implement a binary segmentation algorithm as in [2], i.e. an iterative procedure which looks for a single changepoint for the whole dataset and, if found, iteratively splits the data at the changepoint dealing with the resulting segments separately until no more changes are detected in any segment. This procedure can be matched with either method for a single changepoint detection proposed in [1].
1. The **Bayes Factor method** (BF): a changepoint is found in location θ^* iff $\gamma = \pi(\theta^*) + l_1^* - l_0 > 0$, where θ^* is the location returning the highest marginal log-likelihood, $\pi(\theta^*)$ is the prior probability assigned to that value, l_1^* is the corresponding log-likelihood and l_0 is the log-likelihood under the null hypothesis of no changepoint.
2. The **Posterior Threshold method** (PT): a threshold is chosen and if there are posterior probability

values above the threshold, the highest peak marks the detected changepoint location. For discussion about the choice of the threshold, we refer to [1].

3 A Log-Gaussian Cox Process for earthquake data with changepoints

A Bayesian changepoint model needs prior settings on number and positions of the changes, plus a hierarchical model for the data segments. We look for an unknown number of changes at unknow timepoints. We take a uniform prior for the number $m = 1, \dots, M$ of changepoints and we assume a minimum segment length of d time points in order to avoid unrealistic adjacent changes. Considering that changepoints are looked for sequentially, our prior setting can be written as

$$\begin{aligned} \pi(m) &= (M+1)^{-1} \text{ for } m = 0, \dots, M \\ \pi(\theta_1, \dots, \theta_m | m) &= \pi(\theta_m | \theta_{m-1}, m)\pi(\theta_{m-1} | \theta_{m-2}, m) \dots \pi(\theta_1 | m) \text{ where } \pi(\theta_1 | m) = (T - 2 \times d)^{-1} \end{aligned} \tag{1}$$

The conditional priors for $\theta_2, \dots, \theta_m$ can be computed sequentially as the binary segmentation algorithm proceeds. As for the data segment likelihood, we build a model as follows:

$$Y_{ts} \sim Poi(\lambda_{ts}|C|) \text{ with } \log(\lambda_{ts}) = \beta_0 + \phi_t + \psi_s \tag{2}$$

Here Y is a response vector of cell counts for each cell C in a regular grid admitted to the observation window. To model the parameter λ_{ts} (where t indexes time and s space) we use a spatio-temporal Log-Gaussian Cox Process (LGCP) [3], i.e. the logarithm of the intensity function at every location s is assumed to be a Gaussian field and depend on an intercept $\beta_0 \sim N(0, \sigma_\beta^{-2})$ and on two random effects modelled as Intrinsic Gaussian Markov Random Fields. In particular, $\phi \sim IGMRF(0, \tau_\phi K_\phi)$ is a RW(1) over time, and $\psi \sim IGMRF(0, \tau_\psi K_\psi)$ is a RW in two dimensions on a regular grid [1]. The same hyperprior is taken on the precision parameter $\tau_\phi, \tau_\psi \sim Gamma(1, .00005)$ because the IGMRFs are scaled in order to have the same variance, following [5]. LGCPs constitute a broad and flexible class of point process models whose estimation issues have been recently overcome by gridding data and using GMRF processes. They also allow spatial and temporal dependence to be included in the model. Goodness-of-fit tests for point processes based on interevent distances that are routinely used in point process analysis (see for example [3]), indicate that the LGCP model fits the data well; we can thus proceed to the changepoint analysis.

4 Results and discussion

At this first stage, considering the length of the series we assume there are no more than $M = 4$ changepoints, and we assume $d = 2$. Following the prior setting in (1), we write $\pi(m) = 5^{-1}$ and $\pi(\theta_1|m) = 26^{-1}$. As for the data likelihood, we estimate model (2) which we label as 'spatio-temporal', as well as a 'fixed' model including β_0 only, a 'temporal' model including ϕ and a 'spatial' model including ψ.
As regards the detected changepoint locations (Table 1), some findings should be dealt with carefully since 2012 is very close to the end of the series and other changes are only detected in one model scenario. Overall, we can appreciate the detection of a changepoint in 2008; indeed, after 2008 two major seismic events (in L'Aquila and in the Emilia-Romagna region) shocked Italy. We can see in Figure 2 that the average intensity of the process, i.e. the expected number of events per cell, increased due to the mentioned shocks (left panel). Moreover, the spatial distribution changed: until 2008 earthquakes

were evenly distributed all along the Appennini (central panel); afterwards, we see a clusterisation of the process around the central-east part of Italy (where Emilia-Romagna and L'Aquila are) and the volcanic islands close to Sicily (right panel), while a decrease occurs in the Adriatic sea and south-eastern area. As in several applications, it would be of interest to include extra knowledge (such as covariates or informative priors) in order to improve the reliability of the results. Useful information regards number and sensitivity of the monitoring stations and their evolution over time. The detection of earthquakes is related to the distance from the hypocentre and to the magnitude of the event; it might be of interest to investigate whether a higher density of the process is partially due to an increased ability to record seismic events. Moreover, the depth of the hypocentre may be exploited in order to check if it is negatively correlated to the earthquake magnitude; besides, a changepoint analysis of the depth itself may bring useful knowledge to the interpretation of the phenomenon.

Model	Bayes Factor	Posterior Threshold
fixed	1987-2008-2012	1987-2008-2012
temporal	2012	1991-2001-2012
spatial	—	2008
spatio-temporal	—	2008

Table 1: Detected changepoints. Changepoints coloured in red are the ones detected first.

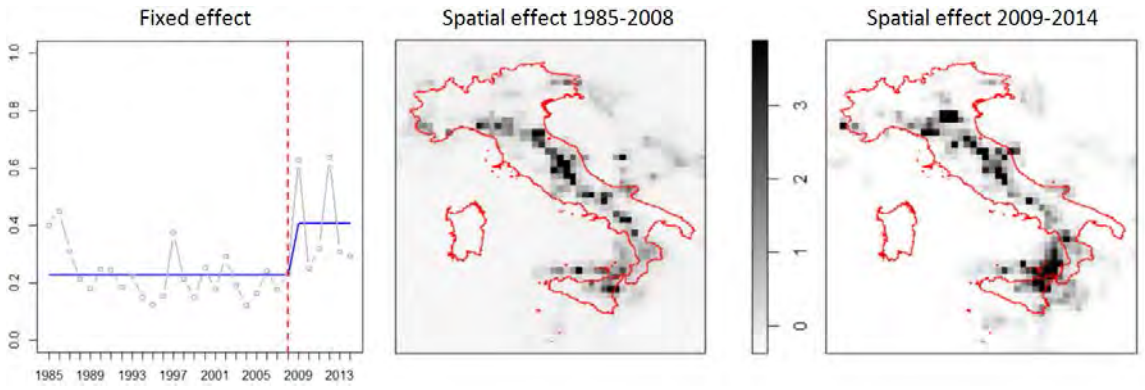


Figure 2: Estimate for the mean intensity function (blue line) and number of events per cell (grey line); spatial effect before and after the main changepoint.

References

[1] Altieri, L., Scott, M., Cocchi, D., Illian, J. (2015). A changepoint analysis on spatio-temporal point processes. *Submitted*

[2] Chen, J. and Gupta, A. K. (2012). Parametric Statistical Change Point Analysis. Birkhauser, Boston

[3] Møller, J., Waagepetersen, R. (2006). Modern statistics for spatial point processes. Aalborg: Department of Mathematical Sciences, Aalborg University. (Research Report Series; No. R-2006-12).

[4] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of The Royal Statistical Society B*, **71**, 319-392

[5] Sørbye, H. S., Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**, 39-51

FANOVA models in rectangular and circular domains

J. Álvarez-Liébana[1] and M. D. Ruiz-Medina[1]

[1] *Department of Statistics and O.R., Univ. of Granada, Granada, SPAIN; mruiz@ugr.es, javialvaliebana@ugr.es*

Abstract. *FANOVA models on rectangles, circular disks and circular sectors are analyzed. Dirichlet boundary conditions are imposed to define the corresponding covariance operators of the Hilbert-valued components of the vector error term. Minimal conditions on the design matrix are imposed to derive a generalized least squares estimator of the Hilbert-valued vector of fixed effects.*

Keywords. *Generalized least-squares functional estimation; FANOVA models; Reproducing kernel Hilbert space; Hilbert-valued multivariate fixed effect model; Linear functional tests.*

1 Introduction.

The functional analysis of variance is implemented, after suitable transformation of the functional data model, in the geometry of the Reproducing Kernel Hilbert Space (RKHS). A finite-dimensional chi-squared hypothesis testing is implemented in terms of vectorial projections for the significance analysis of the functional fixed effects. A simulation study is undertaken to illustrate the performance of the proposed methodology, and the the influence of the functional form of the fixed effect parameters, of the geometry of the domain, and of the truncation order is analyzed.

2 FANOVA on rectangular and circular domains.

2.1 The model.

In the following, H will denote a separable Hilbert space, D a Dirichlet regular bounded open domain, $-\Delta_D$ the negative Laplace operator on D, and $Y(\cdot)$ represents H^n-valued variables. In [4], the following model is introduced:

$$Y(\cdot)=[Y_1(\cdot),\ldots,Y_n(\cdot)]^T=X\beta(\cdot)+\varepsilon(\cdot), \qquad E[Y]=X\beta, \qquad E[\varepsilon]=E[\varepsilon_1(\cdot),\ldots,\varepsilon_n(\cdot)]^T=\overrightarrow{0}, \quad (1)$$

where $X\in\mathbb{R}^{n\times p}$ such that $X^TX=Id_p$, $\beta(\cdot)=[\beta_1(\cdot),\ldots,\beta_p(\cdot)]^T\in H^p$, and $\varepsilon(\cdot)$ is a correlated H^n-valued standard Gaussian random variable, which covariance operator matrix is given by

$$\mathbb{R}_{\varepsilon\varepsilon}=\begin{pmatrix}\mathbb{R}_{\varepsilon_1\varepsilon_1} & \ldots & \ldots & \mathbb{R}_{\varepsilon_1\varepsilon_n}\\ \ldots & \ldots & \ldots & \ldots\\ \mathbb{R}_{\varepsilon_n\varepsilon_1} & \ldots & \ldots & \mathbb{R}_{\varepsilon_n\varepsilon_n}\end{pmatrix}, \text{ with } \mathbb{R}_{\varepsilon_i\varepsilon_j}=E[\varepsilon_i\otimes\varepsilon_j], \ \forall i,j=1,\ldots,n,$$

under assumption that $\mathbb{R}_{\varepsilon_i\varepsilon_i}=f_i(-\Delta_D)$, $\forall i=1,\ldots,n$, is strictly positive and compact self-adjoint, in the trace class on $H=L_0^2(D)$, we have (see [1])

$$\lambda_{ki} \quad = \quad f_i(\lambda_k) \ \text{ eigenvalues of } \mathbb{R}_{\varepsilon_i\varepsilon_i}, \quad \mathbb{R}_{\varepsilon_i\varepsilon_i}\phi_k=\lambda_{ki}\phi_k, \ k\geq 1, \ i=1,\ldots,n, \qquad (2)$$

$$\varepsilon_i = \sum_{k=1}^{\infty} \sqrt{\lambda_{ki}}\eta_{ki}\phi_k, \quad E\left[\eta_{ki}\eta_{pj}\right] = \delta_{k,p}\left((1-\delta_{i,j})\frac{e^{-\frac{|i-j|}{\lambda_{ki}+\lambda_{pj}}}}{\sqrt{\lambda_{ki}\lambda_{pj}}} + \delta_{i,j}\sqrt{\lambda_{ki}\lambda_{pj}}\right), \tag{3}$$

$$\mathbb{R}_{\varepsilon_i\varepsilon_j} = \sum_{k=1}^{\infty}\left(\delta_{i,j}^{*}e^{-\frac{|i-j|}{\lambda_{ki}+\lambda_{kj}}} + \delta_{i,j}\sqrt{\lambda_{ki}\lambda_{kj}}\right)\phi_k\otimes\phi_k, \tag{4}$$

where $\delta_{i,j}^{*} = 1-\delta_{i,j}$, $\{\eta_{ki}\}_{k\geq 1,\ i=1,\ldots,n} \sim N(0,1)$, $\{\Phi_k\}_{k\geq 1}$ and $\{\lambda_k\}_{k\geq 1}$ eigenfunctions and eigenvalues of $-\Delta_D$ respectively, and $\{f_i\}_{i=1\ldots,n}$ are continuous decreasing functions. Given an orthonormal set of eigenfunctions $\{\Phi_k\}_{k\geq 1}$ of H, we denote Φ^{*} as $\Phi^{*}(f) = \left\{\Phi_k^{*}(f)\right\}_{k\geq 1} = \left\{\left(\langle f_1,\phi_k\rangle,\ldots,\langle f_n,\phi_k\rangle\right)^T\right\}_{k\geq 1}$, that represents the projection of f on $\{\Phi_k\}_{k\geq 1}$. The inverse operator Φ is given by $\Phi\left(\left\{f_k^T\right\}_{k\geq 1}\right) = \left(\sum_{k=1}^{\infty} f_{k1}\phi_k,\ldots,\sum_{k=1}^{\infty} f_{kn}\phi_k\right)^T$. Also we get $\Phi^{*}\mathbb{R}_{\varepsilon\varepsilon}\Phi = \{\Lambda_k\}_{k\geq 1}$ and

$$\langle f,g\rangle_{\mathbb{R}_{\varepsilon\varepsilon}^{-1}} = \mathbb{R}_{\varepsilon\varepsilon}^{-1}(f,g) = \sum_{k=1}^{\infty} f_k^T\Lambda_k^{-1}g_k,\ \forall f,\ g\in\mathbb{R}_{\varepsilon\varepsilon}^{1/2}(H^n),\ \Lambda_{kij} = e^{-\frac{|i-j|}{\lambda_{ki}+\lambda_{kj}}}\ (i\neq j),\ \Lambda_{kii} = \lambda_{ki}. \tag{5}$$

2.2 Negative Laplacian operator on Dirichlet regular bounded open domains: rectangle, disk and circular sector.

Explicit formulae of the eigenfunctions and eigenvalues of the negative Laplacian operator on rectangular and circular domains are given in [2] for different type of boundary conditions. Here, Dirichlet boundary conditions are considered. That is,

$$-\Delta_D = -\frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2}, \quad -\Delta_D\phi_k(x) = \lambda_k\phi_k(x)\ \left(x\in D\subseteq\mathbb{R}^2\right), \quad \phi_k(x) = 0\ (x\in\partial D). \tag{6}$$

According to [6], in rectangular domains $D = \prod_{i=1}^{2}[a_i,b_i]$, we get the asymptotics $\lambda_k^{-\gamma} = \mathrm{O}\left(k^{-\frac{2\gamma}{d}}\right)$, $k\to\infty$. From [2],

$$\phi_k(x) = \phi_{k_1}^{(1)}(x_1)\,\phi_{k_2}^{(2)}(x_2),\ \lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad \lambda_{k_i}^{(i)} = \frac{\pi^2(k_i+1)^2}{l_i^2}, \tag{7}$$

$$\phi_{k_i}^{(i)}(x_i) = \sin\left(\frac{\pi(k_i+1)(b_i-x_i)}{l_i}\right),\ \forall x_i\in[a_i,b_i],\ l_i = b_i-a_i,\ 1\leq k_i\leq k,\ \forall i=1,2, \tag{8}$$

where $k=1,\ldots,trunc$, with $trunc$ denoting the truncation parameter, that in the case of rectangle domains is $trunc = TR\times TR$, with TR being one-dimensional truncation order at each coordinate. In the case of $D = \left\{x\in\mathbb{R}^2:\ 0<\|x\|<R\right\}$, its rotation symmetry allows us to define $-\Delta_D$ in polar coordinates as

$$-\Delta_D = -\frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} - \frac{1}{r^2}\frac{\partial^2}{\partial\varphi^2}, \quad \lambda_{kh} = \frac{\alpha_{kh}^2}{R^2}, \tag{9}$$

$$\phi_{kh}(r,\varphi) = J_k(\alpha_{kh}r/R)\left(cos(k\varphi)+sin(k\varphi)\right),\ 0<r<R,\ \varphi\in[0,2\pi], \tag{10}$$

where $k=1,\ldots,trunc$ and α_{kh} are the $trunc_k$ positive roots of $J_k(z)$, $z\in[0,R]$, with $h=1,\ldots,trunc_k$. The following asymptotic formulae hold:

$$j_{kh} = k+\delta_h k^{1/3}+\mathrm{O}\left(k^{-1/3}\right)$$
$$j_{kh} = \pi(h+k/2-1/4)+\mathrm{O}\left(k^{-1}\right)$$

for the zeros of Bessel functions $J_k(z)$ of the first kind and order k on circular domains (see [3]; [5]). In particular, for the circular sector of radius R and angle $\pi\theta$, $\phi_{kh}(r,\varphi) = J_{k/\theta}(\alpha_{kh}r/R)\sin(k\varphi/\theta)$, $0<r<R$, $0<\varphi<\pi\theta$, $\lambda_{kh} = \frac{\alpha_{kh}^2}{R^2}$.

3 Generalized least-squares estimator and FANOVA statistics

From [4], considering the geometry of the RKHS,

$$\|Y-X\beta\|^2_{\mathbb{R}^{-1}_{\varepsilon\varepsilon}} = \mathbb{R}^{-1}_{\varepsilon\varepsilon}(\varepsilon,\varepsilon),\ \widehat{\beta} = \left(\sum_{k=1}^{\infty}\widehat{\beta_{k1}}\phi_k,\ldots,\sum_{k=1}^{\infty}\widehat{\beta_{kp}}\phi_k\right)^T,\ \varepsilon = \Phi\left(\{M_kY_k\}_{k\geq 1}\right), \tag{11}$$

$\widehat{\beta}_k = \left(X^T\Lambda_k^{-1}X\right)^{-1}\Lambda_k^{-1}X^TY_k$, $M_k = Id_n - X\left(X^T\Lambda_k^{-1}X\right)^{-1}X^T\Lambda_k^{-1}$ and $\sum_{k=1}^{\infty} trace\left(X^T\Lambda_k^{-1}X\right)^{-1} < \infty$. For the FANOVA analysis, we consider the transformation of our functional data model (1) by a suitable matrix operator $\mathbf{W}$ satisfying the conditions formulated in [4], in order to ensure almost surely finiteness of the functional components of variance. The residual variance, the total sum of squares and the explained functional variability are respectively given by:

$$\begin{aligned}\widetilde{SSE} &= \sum_{k=1}^{\infty}(M_kW_kY_k)^T\Lambda_k^{-1}M_kW_kY_k, \quad \widetilde{SST} = \sum_{k=1}^{\infty}Y_k^TW_k^T\Lambda_k^{-1}W_kY_k, \\ \Lambda_k &= \Psi_k\Omega(\Lambda_k)\Psi_k^T,\ W_k = \Psi_k\Omega_k(W_k)\Psi_k^T,\ \Omega(\Lambda_k) = diag\left(\omega_1(\Lambda_k),\ldots,\omega_n(\Lambda_k)\right),\ \forall k\geq 1,\end{aligned} \tag{12}$$

and $\widetilde{SSR} = \widetilde{SST} - \widetilde{SSE}$, where $\{\psi_{ki}\}_{k\geq 1,\ i=1,\ldots,n}$ and $\{\omega_i(\Lambda_k)\}_{k\geq 1,\ i=1,\ldots,n}$ are the eigenvectors and eigenvalues of Λ_k respectively, and $\Omega(W_k) = diag(w_{k11},\ldots,w_{knn})$ is given by $w_{kii} = \omega_i(\Lambda_k) + \frac{1}{a_k}$, under $\sum_{k=1}^{\infty}\frac{1}{a_k} < \infty$. We get $a_k = k^2$. Lastly, the significance of functional fixed effects is tested as follows: For for $k = 1,\ldots,trunc_k$,

$$H_{0k} : K_k\beta_k = C_k, \quad C_k = (0,0,\ldots,0)^T \in \mathbb{R}^{(p-1)\times 1},\ \forall k = 1,\ldots,trunc_k, \tag{13}$$

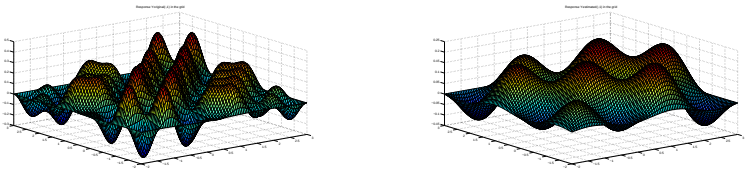
$$T_k = \left(K_k\widehat{\beta}_k - C_k\right)^T\left(K_kX_{\Lambda_k}K_k^T\right)^{-1}\left(K_k\widehat{\beta}_k - C_k\right) \sim \chi^2_{p-1}, \tag{14}$$

$$p-value_k = 1 - P\left(\chi^2_{p-1}\leq T_k\right), \quad K_k = \Phi_k^*K\Phi_k = \begin{pmatrix}1 & -1 & 0 & \ldots & 0\\ 1 & 0 & -1 & \ldots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & 0 & 0 & \ldots & -1\end{pmatrix} \in \mathbb{R}^{(p-1)\times p}.$$

3.1 Results

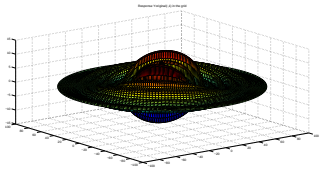
The numerical simulations performed are shown in Figures 1 and 2. We have considered the square-integrable space, $H = L_0^2(D) = \overline{C(D)}^{L^2(\mathbb{R}^2)}$, with compact support. The following empirical least-squares errors have been computed, additionally to the F-statistics for the truncated FANOVA analysis (see Table 3), to illustrate the performance of the proposed rectangular and circular analysis of variance:

$$FMSE_{\beta_s}(\cdot) = \|\beta_s(\cdot) - \widehat{\beta}_s(\cdot)\|^2,\ \forall s = 1,\ldots,p, \quad FMSE_Y(\cdot) = \sum_{i=1}^{n}\frac{\|Y_i(\cdot) - \widehat{Y}_i(\cdot)\|^2}{n}. \tag{15}$$

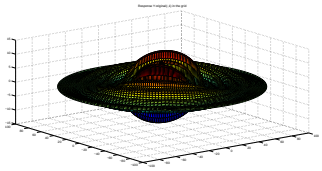


(a) Simulated response with rectangle domain. (b) Estimated response with rectangle domain.

Figure 1: Responses with rectangle domain.



(a) Simulated response with disk domain.



(b) Estimated response with disk domain.

Figure 2: Responses with disk domain.

n	p	TR	$a_1 = a_2$	$b_1 = b_2$	$h_x = h_y$	$\gamma_i \left(\lambda_{ki} = \lambda_k^{-\frac{\gamma_i}{2}}\right)$	$\beta_s(x,y),\ s = 1,\ldots,p$
200	4	4	-2	3	0.05	$1+\frac{i}{n}$	$\beta_s(x,y) = \sin\left(\frac{\pi s x_{b_1}}{l_1}\right)\sin\left(\frac{\pi s y_{b_2}}{l_2}\right)$

Table 1: Parameters in the rectangle domain.

n	p	R	TR	h_R	h_ϕ	C	$\gamma_i \left(\lambda_{ki} = C\lambda_k^{-\frac{\gamma_i}{2}}\right)$	$\beta = \boldsymbol{Comb} * \Phi,\ s = 1,\ldots,p$
200	9	100	1	$\frac{R}{145}$	$\frac{2\pi}{135}$	R^2	$4+\frac{2i}{n}$	$Comb_{sk} = \frac{1}{R}e^{\frac{s+\frac{k}{R}}{n}} + k\cos\left((-1)^k 2\pi\frac{R}{k}\right)$

Table 2: Parameters in the disk domain.

Case	$\|\|FMSE_Y\|\|_\infty$	$\max_{s=1,\ldots,p} \|\|FMSE_{\beta_s}\|\|_\infty$	F statistics
Rectangle	0.0014	0.0015	1.9263
Disk	0.0007	0.0027	$(1.4)10^7$

Table 3: Results.

Acknowledgments. This work has been supported in part by project MTM2012-32674 (co-funded with European Regional Development Funds).

References

[1] Dautray, R. and Lions, J. -L. (1990). *Mathematical Analysis and Numerical Methods for Science and Technology Volume 3: Spectral Theory and Applications*. Springer. New York.

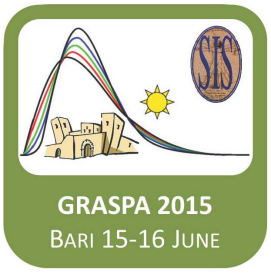
[2] Grebenkov, D. S. and Nguyen, B. -T. (2013). Geometrical structure of Laplacian eigenfunctions. *SIAM Review* **55**, 601–667.

[3] Olver, F. W. J. (1952). Some new asymptotic expansions for Bessel functions of large orders. *Mathematical Proceedings of the Cambridge Philosophical Society* **48**, 414–427.

[4] Ruiz-Medina, M. D. (2014). Functional Analysis of Variance for Hilbert-Valued Multivariate Fixed Effect Models (submitted).

[5] Watson, G. N. (1966). *A treatise on the theory of Bessel functions*. Cambridge University Press. Cambridge.

[6] Weyl, H. (1911). Das asymptotische Verteilungsgesetz der Eigenwerte linearer Differentialgleichungen. *Mathematische Annalen* **71**, 441–469.



Spatio-temporal wind speed predictions for Germany

D. Ambach[1,*] and W. Schmid[1]

[1] *Department of Statistics, European University Viadrina PO Box 1786, 15207 Frankfurt (Oder), Germany; ambach@europa-uni.de, schmid@europa-uni.de* [*]*Corresponding author*

Abstract. *State of the art wind forecasting models, like Numerical Weather Predictions utilize huge amounts of computing time. Some of them have rather low spatio-temporal resolution. Time series prediction model accomplish good results in high temporal settings. Moreover, their consumption of computing capacities is relatively low and return accurate short-term to medium-term forecasts. The recent literature shows increasing interest in the topic of spatial interdependence. This article deals with a spatial and temporal model for wind speed. We describe the temporal model structure independently on spatial correlations. Therefore, seasonality and a huge correlation structure are included. Subsequently, the model is extended and a spatial structure is included. The data set includes ten minute observations of several measurement stations in Eastern Germany. The validation procedure shows that the model is reliable.*

Keywords. *Wind speed; Forecasting; Spatio-temporal; Periodicity; Autocorrelation.*

1 Introduction

The European Wind Energy Association (2014) finds that the cumulative capacity installation for renewables has a total share of 72%. The World Wind Energy Association (2014) indicates that especially the wind energy capacity within Europe increases from 2.4% in 2000 to 14.1% in the year. This indicates the growing relevance of renewable energies for the European market. In particular the European Wind Energy Association (2014) recognises the need for a European Energy Union and the importance of wind energy which is displayed in different wind energy scenarios for 2020. In Europe the wind energy installation is expected to meet about one sixth of the total energy consumption (e.g., European Wind Energy Association, 2014). The Berkhout et al. (2013) point out that about 8% of the German electricity mix is based on wind energy.

Models for the description and prediction of wind speed are of outstanding importance. However, the accuracy of those models differ severely. The characteristics of wind speed and wind energy makes it challenging to forecast the underlying processes precisely. The uncertainty in the data set which remains, maintain the necessity to develop and optimise forecasting methods.

Beside models like Numerical Weather Predictions (NWP) and Artificial Neural Networks (ANN),

there exists a huge scope of time series approaches. A new field of research are hybrid models. They combine two or more types of wind forecasting approaches. We combine temporal and spatial modelling. Damousis et al. (2004) combine a spatial model with a genetic algorithm for model fitting and forecasting. Han and Chang (2010) describe a simulation study to analyse the impact of spatial and temporal correlation on wind power forecasting accuracy. Moreover, Benth and Šaltyte (2011) evolve a spatial and temporal model. Primarily, they model the temporal structure by using a periodic autoregressive moving average model (ARMA) and a heteroscedastic variance. Accordingly, they use Gaussian random fields to cover the spatial dependence. Hering and Genton (2010) as well as Zhu et al. (2014) cover spatial dependence structure by regime switching models. Following this research, Díaz et al. (2014) comment on the importance of spatial composition according to the spread and wake effect of wind generators over a small geographical areas. Another important aspect of spatial wind modelling is related to the characterization of wind resources at specific regions where sufficient information is not available. In this situation, kriging can be applied. Thus, spatio-temporal models are useful indicators to perform wind energy potential assessments at sites without measurements (e.g Jung and Broadwater, 2014).

The article is structured in the following way. Section 2 describes the data set. Besides, we analyse the spatial structure. The novel spatio-temporal model is introduced in Section 3. Section 4 provides a small outlook of our results.

2 Wind speed data in Eastern Germany

In Figure 1 the measurement stations of the considered wind speed data are shown. They are located in Brandenburg and Berlin. This article focuses on Eastern Germany according to the homogeneity of this region. This area is rural and plain and perfect for wind parks. The data is provided by the "Deutscher Wetterdienst" (DWD) and reaches from January 2009 to December 2011. For model fitting, a time span of one years is used and the remaining months (January 2011 to December 2011) are used for out-of-sample forecasts. The wind speed $(W_t)_{t\in\{1,\cdots,T\}}$ is measured in m/s in a 10-minute interval.

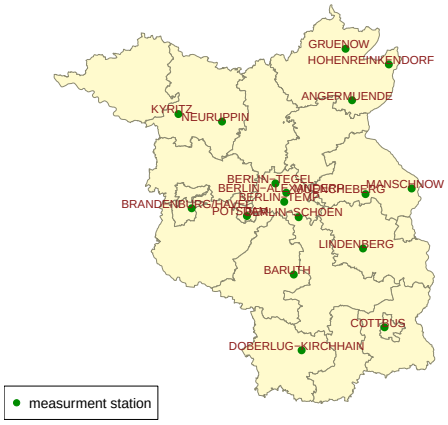


Figure 1: Meterological measurement stations in Berlin and Brandenburg which provide 10min data

3 A spatio-temporal wind speed model

Univariate time series models for wind speed modelling have been considered by Ambach and Schmid (2015). They are based on a periodic temporal model structure. Nevertheless, such models are unable to predict the wind speed at stations where no measurement is observed. Therefore, spatio-temporal models provide a beneficial contribution. The here described methodology, is an extension of Ambach and Croonenbroeck (2015) and is able to capture the spatial and temporal auto-correlation structure of the wind speed data and the cross-correlation with other variables.

We consider a periodic space-time autoregressive model with external regressors for the wind speed (e.g. Cressie and Wikle, 2011). The vector of observations at time t and all locations is denoted by $\boldsymbol{W}_t = (W_{1t}, ..., W_{nt})$. Let $\{W_{st} : s = 1, ...n; t = 1, ..., T\}$ denote the collection of data. Therefore, we obtain the following model equation

$$\boldsymbol{W}_t = \boldsymbol{\alpha}_t + \sum_{j=1}^{p} \boldsymbol{\phi}_{tj} \boldsymbol{W}_{t-j} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\delta} \boldsymbol{v_t} + \boldsymbol{\epsilon}_t, \tag{1}$$

$$\boldsymbol{\alpha}_t = \boldsymbol{\vartheta_0} + \boldsymbol{U}_1 \sum_{i=1}^{K} \boldsymbol{\vartheta}_i B_i^{\boldsymbol{s}}(t), \tag{2}$$

$$\boldsymbol{\phi}_{tj} = \boldsymbol{U}_2 \boldsymbol{\phi}_{0j} + \boldsymbol{U}_3 \sum_{i=1}^{K} \boldsymbol{\phi}_{ij} B_i^{\boldsymbol{s}}(t), \tag{3}$$

where $\{\boldsymbol{\epsilon_t}\} \sim N_n\big(\boldsymbol{0}, \boldsymbol{\sigma_\epsilon^2 I}\big)$ is an $n \times 1$ column vector. The measurement errors $\boldsymbol{\epsilon_t}$ follow a Gaussian white noise in space and time. Furthermore, $\boldsymbol{v_t}$ is an $n \times 1$ vector which contains spatial random effects and $\boldsymbol{X_t}$ is a matrix of external variables (e.g. Finazzi et al., 2013). These effects are time independent, but provide the following spatial correlation structure $\boldsymbol{v_t} \sim N_n\big(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{H}, \boldsymbol{\theta})\big)$. The spatial covariance matrix $\boldsymbol{\Sigma}$ is determined by $\boldsymbol{H}$, which is a matrix of pairwise geographic distances and a spatial covariance function $(C(s_h, s_l, \theta))_{h,l=1,...n}$. Moreover, $\boldsymbol{\vartheta_0}$ is an $n \times 1$ intercept vector and $\boldsymbol{\vartheta}_{i_1}$ are $n \times 1$ periodic coefficient vectors. $\boldsymbol{\phi}_{0j}$ is an $n \times n$ parameter matrix of autoregressive parameters for lag j and $\boldsymbol{\phi}_{j_1 j}$ are $n \times n$ parameter matrices of periodic AR parameters for lag $(j)_{j \in \mathbb{N}}$. Furthermore, $\boldsymbol{U}_1$, $\boldsymbol{U}_1$ and $\boldsymbol{U}_3$ are known spatial weight matrices of dimension $n \times n$. De Boor (1978) and Eilers and Marx (1996) define the fundamentals for the periodic B-spline basis functions which is given by $B_i^{\boldsymbol{s}}(t)$. Therefore, it is important to define the set of equidistant knots κ. The daily periodicity is $s = 144$. Cubic B-splines are an attractive approach, because they are twice continuously differentiable.

In this study we use a model which is able to perform predictions for a certain regions. Moreover, it is able to capture the temporal structure of our wind speed data set. We include the aforementioned periodicities as well as other known regressors. Especially, the terrains roughness, natural vegetal cover, and meteorological variables. The meteorological variables are not deterministic, but we are able to use the temporal lagged information. Hence, we are able to include linear mixed effects to introduce more spatial and temporal structure.

4 Outlook

The main objective of this article is to provide a model for wind speed which is easy to use and provide reliable prediction in both space and time. We recommend a simple periodic space-time autoregressive model with external regressors model which seems powerful enough to describe the wind dynamics in

both dimensions. The model contains periodic B-spline functions and space and time autoregressive components. Although, the wind speed is a noisy meteorological variable, which makes it hard to model, the described model provides a useful procedure to predict the spatio-temporal wind speed structure.

References

Ambach D, Croonenbroeck C (2015) Using the lasso method for space-time short-term wind speed predictions. arXiv preprint arXiv:150106406

Ambach D, Schmid W (2015) Periodic and long range dependent models for high frequency wind speed data. Energy 82(0):277 – 293

Benth JŠ, Šaltyte L (2011) Spatial–temporal model for wind speed in lithuania. Journal of Applied Statistics 38(6):1151–1168

Berkhout V, Faulstich S, Görg P, Hahn B, Linke K, Neuschäfer M, Pfaffel S, Rafik K, Rohrig K, Rothkegel R, Zieße M (2013) Wind energy report germany 2013. Fraunhofer-Institut für Windenergie und Energiesystemtechnik–IWES–Kassel

Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. John Wiley & Sons

Damousis I, Alexiadis MC, Theocharis JB, Dokopoulos PS (2004) A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation. In: IEEE Transactions on Energy Conversion, vol 19, pp 352 – 361

Díaz G, Casielles PG, Coto J (2014) Simulation of spatially correlated wind power in small geographic areas – sampling methods and evaluation. Electrical Power and Energy Systems 63:513 – 522

De Boor C (1978) A practical guide to splines. Mathematics of Computation

Eilers PH, Marx BD (1996) Flexible smoothing with b-splines and penalties. Statistical science pp 89–102

European Wind Energy Association EWEA (2014) Wind in power, 2013 european statistics, 2014. EWEA: Brussels

Finazzi F, Scott EM, Fassò A (2013) A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of scottish air quality data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 62(2):287–308

Han Y, Chang L (2010) A study of the reduction of the regional aggregated wind power forecast error by spatial smoothing effects in the maritimes canada. In: 2nd IEEE International Symposium on Power Electronics for Distributed Generation Systems, pp 942 – 947

Hering AS, Genton MG (2010) Powering up with space-time wind forecasting. Journal of the American Statistical Association 105(489):92–104

Jung J, Broadwater RP (2014) Current status and future advances for wind speed and power forecasting. Renewable and Sustainable Energy Reviews 31:762–777

World Wind Energy Association WWEA (2014) 2014 half year report. WWEA: Bonn

Zhu X, Genton MG, Gu Y, Xie L (2014) Space-time wind speed forecasting for improved power system dispatch. Test 23(1):1–25

Mixture of experts for sequential PM10 forecasting in Normandy (France)

B. Auder [1], J. M. Poggi [2,*] and B. Portier [3]

[1] *Univ. Paris-Sud Orsay, benjamin.auder@math.u-psud.fr*
[2] *Univ. Paris-Sud Orsay et Univ. Paris Descartes, jean-michel.poggi@parisdescartes.fr*
[3] *Normandie Université, INSA Rouen, bruno.portier@insa-rouen.fr*

[*] *Corresponding author*

Abstract. *Within the framework of air quality monitoring in Normandy, we experiment the methods of sequential aggregation for forecasting concentrations of PM10 of the next day. Besides the field of application and the adaptation to the special context of the work of the forecaster, the main originality of this contribution is that the set of experts contains at the same time statistical models built by means of various methods and different sets of predictors, as well as experts which are deterministic chemical models of prediction modeling pollution, weather and atmosphere.*

Numerical results on recent data from April 2013 until March 2014, on three monitoring stations, illustrate and compare various methods of aggregation. The obtained results show that such a strategy improves clearly the performances of the best expert both in errors and in alerts and reaches an unbiased observed-forecasted scatterplot, difficult to obtain by usual methods.

Keywords. *Air quality; Forecasting; Mixture of experts; PM10; Sequential prediction.*

1 Introduction

In Normandy (France), Air Normand together with Air COM monitor air quality. During a recent research project between academia and air quality agency (see Poggi, Portier 2011), the statistical forecasting of PM10 has been considered, with the aim of improving warning procedures. It led to the development of operational procedures allowing to forecast the daily average of the PM10 for the current day and for the next day on various horizons of forecast integrating the meteorological information and the model outputs statistics.
More generally, Air Normand has various operational tools for the analysis of episodes and for the interpretation of measures, in view of decisions. However these complementary tools, statistical or deterministic models, local or global, often supplying different forecasts especially because of the various space and time resolutions considered.
In this paper, we evaluate the interest of using sequential aggregation or mixing of experts to develop decision-making tools for the forecasters of Air Normand.

In the context of sequential prediction, experts make predictions at each time instance, and the forecaster determines, step by step, the future values of an observed time series. To build his prediction, it/he has to combine before each instant the forecasts of a finite set of experts. To do so,

adopting the deterministic and robust view of the literature of the prediction of individual sequences, three basic references can be highlighted: the survey paper by Clemen (1989), the book of Cesa-Bianchi and Lugosi (2006) and the paper by Stoltz (2010) in French.

In the application framework at hand, empirical studies are particularly valuable and we can mention some studies. In the area of climate Monteleoni et al. (2011), in the field of the air quality Mallet (2010), Mallet et al. (2009), the use of the quantile prediction of the number of daily calls in a call center Biau et al. (2011) and finally the prediction of electricity consumption Devaine et al. (2013). These studies focus on the rules of aggregation of a set of experts and examine how to weight and combine these experts.

Besides the field of application and the adaptation to the special context of the work of the forecaster, the main originality of this work is that the set of experts contains together:
- experts coming from statistical models constructed using different methods and different sets of predictors;
- experts defined by deterministic models of physicochemical prediction modeling pollution, weather and atmosphere. The models are of similar nature but of different spatial and time resolutions with or without statistical adaptation;
- and finally references such as persistence, as usual.

The aforementioned studies combine "homogeneous" methods: only statistical methods or only deterministic ones. Sequential prediction allows mixing several models built on very different assumptions in a unified approach that does not require any prior knowledge about the internal way to use for each expert to generate predictions. It is therefore particularly suitable for our application.
Note that the recent reference Gaillard et al. (2014) is also interested in the design of new experts to be included in the combination, suggesting potential to enrich the basic core of the initial experts, using resampling methods.

2 Data, basic models and aggregation methods

The study period extends from April 3, 2013 to March 18, 2014 (351 days). We have measures of the daily average concentration of PM10 (including the volatile fraction) and in addition the forecasts of the day for the day after of the daily average concentration of PM10 coming from 9 different prediction models. We consider two urban background stations in the Air Normand network HRI (Le Havre) and PQV (Rouen) and an urban background station in the Air COM network LIS (Lisieux).

Numerous forecasting models of different nature are available:
- four statistical models (see Poggi, Portier (2011)): a mixture of regression models with two classes, two linear models (one fitted on slightly polluted days and the other one, on polluted days), and a non-linear additive model;
- three numerical models (Esmeralda and two PREV'AIR models at different spatial resolutions (see `http://www2.prevair.org/` and `http://www.esmeralda-web.fr/`);
- two deterministic models with statistical adaptation (Esmeralda and PREV'AIR);
- the persistence model.

Among the methods used for sequential aggregation strategy, we can distinguish two subsets of different nature.
First, the methods, starting from an initial weighting between the experts and initial performance, are changing the weights adaptively updating the weights at each step. In this category, we will focus on the exponential weight method, called **EWA (Exponential Weighted Average)**.
The second category consists of methods that optimize at every step a global criterion on the history of measurements and expert predictions. Of course, the past can be restricted to a window or the observations can be weighted to emphasize recent ones without omitting those of the past however

distant. In this category, we will focus on the minimization of a quadratic criterion with a quadratic penalty on the coefficients of the mixture, which regularize them (it is the **ridge regression (RR)** framework) or with a l_1-penalty which tend to cancel small coefficients (it is the lasso regression framework).

3 The results

The three main conclusions about the value of the strategy of the mixture of experts are as follows.

First, the comparative **performance** is stable across the three considered stations and the contribution of the combination is clear. Qualitatively, the RR method of mixture of experts provides unbiased predicted-observed scatterplots to be compared with the one of best expert and the EWA mixture (see Figure 1). For quantitative performance: RR (Ridge Regression) is dominant for alerts (see the Threat Score TS) and EWA (Exponential Weighted Average) is the best for the RMSE (Root Mean Square Error). In addition, note that the two methods outperform the best expert.

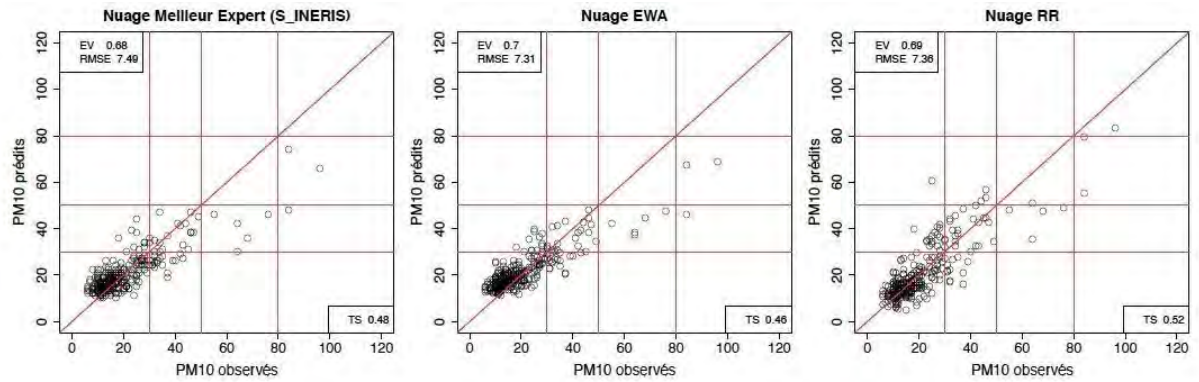


Figure 1: Scatterplots predicted-observed of the mixtures of experts for station HRI (Rouen). Best expert on the left, EWA in the middle and RR on the right.

Even if EWA is not the best one, it deserves to be considered in the future, especially because history will increase and we will more suitably chose the time window. So, they have to be considered as RR methods for deeper investigations in the next step towards operational implementation.

Second, **cooperation between statistical and deterministic** models is helpful, and if two models are often associated with higher weights, all of the models are involved. Indeed, for all the stations, 3 families of methods (deterministic, statistical Air Normand, other statistics) are useful. The sums of the absolute values of weight per family range between 0.2 and 0.4 for EWA and between 0.5 and 3 (and 2 fast enough) to RR. Finally note that the persistence has a small but positive and useful contribution.

Third, if the most useful **methods** are few, no basic method is to depart since an interpretation of the unbiased forecasts of RR and good performance balanced between false alarms and missed alarms, is twofold: first, for each day some methods overestimate and others underestimate (and of course these subsets change with time) and secondly RR (as well as lasso) is best able (with respect to EWA type methods) to take advantage of this situation because it is not constrained by the convexity of weights.

Acknowledgments. This work is the result of a research collaboration between, on the one hand, for the applied side, Air Normand (`http://www.airnormand.fr`) and secondly, for the academic side, the University of Orsay and INSA Rouen. We thank Véronique Delmas, Head of Air Normand, for the problem, the data and for supporting the statistical study. We also thank Pierre Gaillard, Yannig Goude and Michel Bobbia for fruitful discussions.

References

[1] Biau, G., Patra, B. (2010), Sequential quantile prediction of time series. IEEE Transactions on information Theory 57(3), 1664–1674.

[2] Cesa-Bianchi, N., Lugosi, G. (2006), Prediction, Learning, and Games, Cambridge University Press.

[3] Chaloulakou A., Saisana M., Spyrellis N. (2003), Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens, The Science of the Total Environment 313, 1-13.

[4] Clemen, R.T. (1989), Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5(4), 559–583.

[5] Devaine, M., Gaillard, P., Goude, Y., Stoltz, G. (2013), Forecasting electricity consumption by aggregating specialized experts, Machine Learning 90(2), 231–260.

[6] Dietterich T.G. (2000), Ensemble methods in machine learning. In Multiple classifier systems, 1-15, Springer Berlin Heidelberg.

[7] Freund, Y., Schapire, R.E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

[8] Gaillard, P., Stoltz, G., van Erven, T. (2014), A second-order bound with excess losses. ArXiv:1402.2044.

[9] Gaillard, P., Goude, Y. (2015), Forecasting electricity consumption by aggregating experts; how to design a good set of experts. Book chapter to appear in Lecture Notes in Statistics, Springer.

[10] Mallet, V., Stoltz, G., Mauricette, B. (2009), Ozone ensemble forecast with machine learning algorithms. Journal of Geophysical Research 114(D05307).

[11] Mallet, V. (2010), Ensemble forecast of analyses. Coupling data assimilation and sequential aggregation. Journal of Geophysical Research 115(D24303).

[12] Monteleoni, C., Schmidt, G.A., Saroha, S., Asplund, E. (2011), Tracking climate models. Statistical Analysis and Data Mining 4(4), 372–392.

[13] Poggi J.-M., Portier B. (2011), PM10 forecasting using clusterwise regression, Atmospheric Environment, 45(38), 7005-7014.

[14] Siwek K., Osowski S., Garanty K., and Osowski S. (2009), Ensemble of predictors for forecasting the PM10 pollution, Proceedings of VXV International Symposium on Theoretical Electrical Engineering (ISTET), pp. 318-322.

[15] Stoltz G. (2010), Agrégation séquentielle de prédicteurs: méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique, Journal de la Société Française de Statistique, 151(2):66-106.

Evolutionary Polynomial Regression application for missing data handling in meteo-climatic gauging stations

E. Barca[1], L. Berardi[2,*], D. B. Laucelli[2], G. Passarella[1], O. Giustolisi[2]

[1] *Water Research Institute of the National Research Council, Department of Bari, Viale F. De Blasio, 5 70123 Bari, Italy; emanuele.barca@ba.irsa.cnr.it, giuseppe.passarella@ba.irsa.cnr.it*
[2] *Dept. of Civil Engineering and Architecture, Technical University of Bari, Via E. Orabona 4, 70125 Bari, Italy; luigi.berardi@poliba.it, daniele.laucelli@ poliba.it, orazio.giustolisi@poliba.it*
[*] *Corresponding author*

Abstract. *One of the most often encountered modelling problems is that of handling missing data, i.e. the problem of intermediate data gaps, where data/observations before and after the missing observations are available. The gaps in data represent discontinuities, which can pose difficulties both for model construction and model application phases. Evolutionary Polynomial Regression (EPR-MOGA) is a data-driven hybrid technique, which combines the effectiveness of genetic programming with the numerical regression for developing simple and easily interpretable mathematical model expressions. Evolutionary Polynomial Regression takes advantage of the evolutionary computing approach that allows the construction of several model expressions based on training data and least squares methodology to estimate numerical parameters/coefficients. These models can then be verified on a test set and gaps can be in-filled in test datasets by using one selected model. Because of the pseudo-polynomial formulations achievable by EPR-MOGA, it requires fewer numbers of parameters to be estimated, which in turn requires shorter time series for training. Another advantage of the EPR-MOGA approach is the ability to choose objective functions pertaining accuracy and parsimony. In the present work, an application of EPR-MOGA is shown on some sites belonging to the Apulian meteo-climatic monitoring network.*

Keywords. *Evolutionary Polynomial Regression; Missing Data Handling; Environmental Monitoring Networks.*

1 Introduction

In the framework of missing data handling, the need arose for methodologies powerful in addressing the issue and more intuitive in their estimation mechanism, particularly in dealing with variables having a well-known space-time structure such as rainfall and temperature. In the present work, a first attempt to deal with the missing data issue via Evolutionary Polynomial Regression (EPR-MOGA) is shown. EPR-MOGA is a data-modelling hybrid technique, which combines the effectiveness of genetic programming with numerical regression for developing simple and easily interpretable mathematical model expressions. Features that make EPR-MOGA paradigm potentially useful for such applications stem from the reduced number of parameters to tune, which in turn requires shorter time series for model training, and the possibility of building non-linear relationships among input-output data, thus going beyond the linear hypothesis underpinning classical geostatistical approaches.

1 Materials and Method

1.1 Evolutionary Polynomial Regression

Evolutionary Polynomial Regression (EPR) is a data-driven hybrid technique, which combines the

effectiveness of genetic programming with numerical regression for developing simple and easily interpretable mathematical model expressions ([2]). The EPR approach overcomes some drawbacks of other modelling approaches, such as physically based models and black-box data-driven models. The former can be difficult to be constructed due to the underlying mechanisms that may be not always fully understood, or to the need of many data, sometimes difficult to be measured on field. The latter, as for example artificial neural networks, are very effective in reproducing whatever database related to some observed phenomenon, but bring with them some overwhelming problems, like the model structure identification, the over-fitting to training data, and the inability to exploit physical insight about the phenomenon at stake. The EPR can overcome these problems by means of an explicit model expression for the system under observation. EPR-MOGA can be defined as a non-linear global stepwise regression for symbolic data modelling. EPR generalizes the original stepwise regression of [1, 3] by considering non-linear model structures (i.e., pseudo-polynomials) although they are linear with respect to regression parameters. One of the general model structures that EPR-MOGA can manage is reported in Eq. (1):

$$\mathbf{Y} = a_0 + \sum_{j=1}^{\mathrm{m}} a_j \cdot \left(\mathbf{X}_1\right)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot \left(\mathbf{X}_\mathrm{k}\right)^{\mathbf{ES}(j,\mathrm{k})} \cdot f\left(\left(\mathbf{X}_1\right)^{\mathbf{ES}(j,\mathrm{k}+1)} \cdot \ldots \cdot \left(\mathbf{X}_\mathrm{k}\right)^{\mathbf{ES}(j,2\mathrm{k})}\right) \quad (1)$$

where m is the number of pseudo-polynomial terms, a_j are numerical parameters (coefficients) to be estimated, $\boldsymbol{X}_i$ are candidate explanatory variables, $\mathbf{ES}(j,z)$ (with $z = 1,..., 2k$) are the exponents selected from a user-defined set of candidate values (which should include 0), f is a user-selected function (it can be also "no function" resulting into terms obtained by combining input variables). Model parameters are computed from data by solving a linear inverse problem in order to guarantee a two-ways (i.e., unique) relationship between each model structure and its parameters ([2]). From a regressive standpoint, EPR may produce a non-linear mapping of data (like that achievable by Artificial Neural Networks although with few constants. These features, in turn, help avoiding over-fitting to training data thus improving generalization of resulting models. Furthermore, due to the search for model structure, EPR does not require a prior rigid selection of mathematical expressions and number of parameters. Such a flexible coding of mathematical expressions permits to explore the space of the models as the combinatorial space of exponents in Eq. (1). Model search is cast as the solution of a multi–objective optimization problem where fitting to observations (i.e. model accuracy) is maximized while minimizing the complexity of resulting model expressions. Such search exploits the OPTIMized Multi-Objective Genetic Algorithm (OPTIMOGA, [4]) and give rise to a Pareto set of model expressions whose increasing complexity in terms of input variables (i.e. with non-null exponent) and/or number of additional terms, is justified only against an increased fitting performance (i.e. Coefficient of Determination). Due to these features, EPR-MOGA allows to select among optimal models according to the need of the user (e.g. selected model as a trade-off between accuracy and complexity). Additionally, the models can be selected according to the available physical insight about the problem at stake (e.g., recognizing the presence of some known relationship into the explicit formulation of EPR model); conversely, EPR-MOGA can help in discovery some new relationships coming out from the observed data. The EPR-MOGA is available as an add-in function for Excel (Microsoft-Office®) at `www.hydroinformatics.it`.

1.2 Study area, monitoring network and rainfall time series

The proposed method has been applied to the monthly total rainfall time series originating from 81 stations irregularly positioned within the Apulia Region (South-Eastern Italy) (Figure 1). All gauging stations belong to the meteorological monitoring network of the Hydrographic Services of Land Protection Department of the Apulia Region. The time series range from January 1931 to December, 2010. The elevation of each station ranges from 2.00 m a.s.l., (Manfredonia station) to 954.00 m a.s.l. (Pescopagano station) and the average distance between the monitoring stations is around 120 km with a standard deviation of 26 km.

2 Results and Discussion

2.1 Using EPR in missing data reconstruction

In this paper, EPR is used to infill artificially created gaps in rainfall monthly data for the measurement gauge of Canosa ($P^{Can}(t)$), using the observed rainfall monthly data of Cerignola(P^{Cer}) and Andria (P^{And}) gauges. Available monthly rainfall data cover the period from January 1926 to December 2004, without gaps in data for the three rainfall gauges. Data until December 1990 were used as training data to develop EPR-MOGA models. Data from January 1991 to December 2004 where used as test data assuming hypothetic randomly distributed gaps in data, as reported in Table 1. In the case study, the available data has been considered as time series. Thus the inputs used for the estimation of PCan(t) include rainfall monthly data up to 4 month before the time t (e.g., t-1, t-2, t-3 and t-4) (i.e. P^{Cer}, P^{And} and P^{Can}). The set of candidate exponents is [-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2], the maximum number of allowed polynomial terms is 3 and the presence of a bias term is admitted.

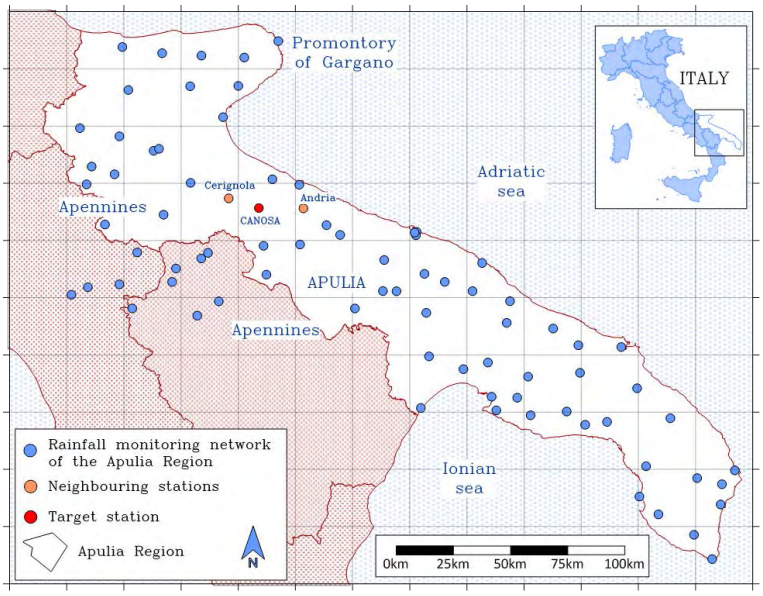


Figure 1. Study area and monitoring sites.

Length of the gap in months	Number of gaps
1	20
2	1
3	1

Table 1. Artificial gaps in the test set.

The EPR-MOGA searching procedure returned 8 models of different complexity, in which the simplest one requires the presence of the only $P^{Cer}(t)$ rainfall monthly value (with a CoD = 0.76), thus indicating that this is the most important input to estimate $P^{Can}(t)$. With the increasing of EPR models accuracy (maximum CoD = 0.82) the complexity also increase, and more data has been selected by the procedure. As a trade-off between accuracy and complexity the following model has been chosen for the present case study:

$$P^{Can}(t) = 0.91\sqrt{P^{And}(t)P^{Cer}(t)} + 4.135$$

This model is featured by a CoD = 0.813, showing the presence of the only rainfall monthly values $P^{Cer}(t)$ and $P^{And}(t)$. Note that rainfall data from previous months are selected only for the most complex models, whose increased number of terms and input variables does not result into an increased accuracy. Accordingly, they are not considered really influent for the estimation of $P^{Can}(t)$. Figure 2 shows the comparison among real and estimated values, while Table 2 shows some error indicators.

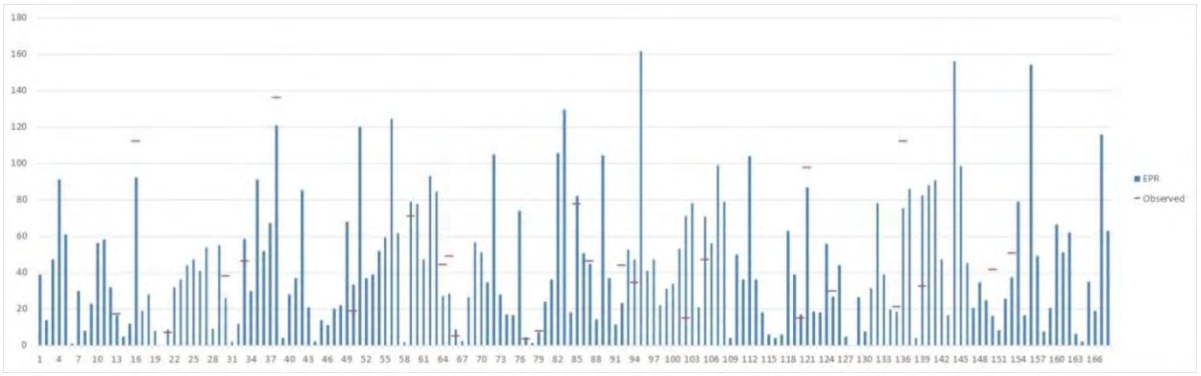


Figure 2. Comparison among real and estimated values of $P^{Can}(t)$.

Maximum Error	56.00 mm
Minimum Error	0.27 mm
Average Error	14.38 mm
Standard deviation of errors	14.44 mm

Table 2. Statistics of the fictitious gaps in the test set.

1.1 Possible future applications of EPR

In the present study, EPR-MOGA has hinted at possible application for identification of multiple correlations among rainfall gauges on a relatively wide territory. Having more time series from a range of gauge station can allow an analysis on how monthly rainfall are distributed and correlated in a (possibly) wide period of time, without excessive data-(pre)processing, but just considering the produced EPR-MOGA models and selected inputs. Considering the analysis of the single gauge station, the availability of different climatic variables (e.g., temperatures, day-night temperature range, wind speed, etc.) can allow the possible correlation with rainfall, aggregated both at monthly and daily scale, eventually allowing missing data reconstruction without resorting to other gauge station records. Furthermore, the obtained result is interesting because proposes a dependence type between the monitoring target station (Canosa) and the neighbouring stations (Andria and Cerignola), different from the linear one commonly used. This approach actually suggests a different paradigm with respect the geostatistical one. In addition, more complex models using an increasing number of neighbouring stations need to be investigated possibly resulting into more reliable filling of missing data.

References

[1] Draper, N.R., Smith, H. (1998) *Applied Regression Analysis*. John Wiley & Sons, New York, 1998
[2] Giustolisi, O., Savic, D.A. (2006).A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinformatics*, **8**, 207-222.
[3] Giustolisi, O., Savic, D.A. (2009).Advances in data-driven analyses and modelling using EPR-MOGA. *J. Hydroinformatics*, **11**, 225-236.
[4] Laucelli, D., Giustolisi, O. (2011) Scour depth modelling by a multi-objective evolutionary paradigm. *Environmental Modelling & Software*, **26**, 498-509.

Data-driven and multi-approach sampling scheme optimization: the Alimini Lakes aquifer case

E. Barca[1,*], M.C. Caputo[1], L. De Carlo[1], R. Masciale[1], G. Passarella[1]

[1] *Water Research Institute of the National Research Council, Department of Bari, Viale F. De Blasio, 5 70123 Bari, Italy; emanuele.barca@ba.irsa.cnr.it; maria.caputo@ba.irsa.cnr.it; lorenzo.decarlo@ba.irsa.cnr.it; rita.masciale@ba.irsa.cnr.it; giuseppe.passarella@ba.irsa.cnr.it.*

[*] *Corresponding author*

Abstract. *Due to the high wells drilling cost, monitoring sites are usually selected among existing wells; nevertheless, the resulting monitoring network must assure a good assessment of the main characteristics of the considered aquifer. Groundwater managers, need to find a good balance between two conflicting objectives: maximizing monitoring information and minimizing costs. In this paper, a couple of groundwater monitoring optimization methods are presented, related to the local shallow aquifer of the Alimini Lakes, located in Apulia (South-Eastern Italy) where a large number of existing wells have been pinpointed and the need of optimally reducing exists. The proposed methods differ each other for the required amount of prior information. The first proposed method, namely Greedy Deletion, just requires the geographical position of the available sites, while the second, the Simulated Annealing, also requires the knowledge of the spatial law of the considered phenomenon. The managerial need was to halve the number of monitoring sites minimizing the information loss.*

Keywords. *Monitoring networks; Shallow aquifers; Greedy deletion; Spatial simulated annealing.*

1 Introduction

Groundwater monitoring is generally rather expensive due to the wells drilling costs. Usually, monitoring networks (MNs) are designed selecting those most capable of representing the groundwater status among a wide number of wells. Nevertheless, the available wells are often irregularly spread and an unwise selection of them may cause a biased understanding of the monitored water body (Barca et al., 2015). However, a wise selection of wells is strongly driven by the a priori knowledge of the considered water body. Different approaches can be found in scientific literature, related to the OMNR, whose application is driven by the available information (Nunes et al., 2002; Wu, 2004). In general, the OMNR is an optimisation problem that is solvable through the quantitative formulation of one or more Objective Functions (OFs). The choice of the OF is strongly dependent on the available information. Nevertheless, this information is not always available simultaneously and constrains the choice of the optimization methodology driving the OF selection (Barca et al., 2015). In practice, when the spatial behaviour of the monitored parameter is known, a model-based method can be used and the OF will be strongly dependent on the variable. Conversely, when the a priori knowledge is poor, only a design-based approach applies, since these latter methods exploit geometric characteristics as OF. In this paper, two OMNR methods are presented belonging to model-based and design-based categories, respectively. In particular, the GD and the SSA methods are applied to the optimal downsize of the Alimini Lakes groundwater monitoring network, initially made of 85 wells, covering a planar area of about 80 km2. Theoretical and applicative issues are reported referred to halving the original network.

2 Materials and Methods

Optimization methods

The OMNR issue can be brought back to a combinatorial problem of extracting a subset of cardinality k with some specified properties (i.e., number of locations to be removed), from a set of cardinality N (i.e., the initial network size) (Barca et al., 2015). The exhaustive exploration of the whole solution space S^N is almost always unfeasible because it is too computationally demanding. Consequently, the use of an optimization heuristics capable of reducing the solution space becomes necessary in order to simplify the issue. In practice, given the general monitoring aim and the a priori available information, a quantitative criterion

$$\phi(S): S^N \rightarrow R^+ \tag{1}$$

where $\phi(S)$ is the OF, must be defined which automatically leads to a specific optimization method category (model/design-based). In general, the stages of an optimization method can be summarized as follows, independently from its category:

1. a starting reduced configuration S_0 is defined;
2. a new candidate configuration S_i is generated by means of a sequential neighbouring search;
3. a decision rule checks each generated configuration and selects the optimal transient solution S^*;
4. a stop criterion states the convergence to the optimal solution, S^O.

In this paper, the GD method and the SSA method belonging to the model-based and design-based categories, respectively, are presented and applied.

Greedy Deletion

The GD method is a greedy heuristic for the optimal reduction of the monitoring network. The initial configuration S_0 is made up of all the locations of the original network. This method substantially operates three nested loops. The outer loop is governed by the problem size, namely the k locations to be removed. The second intermediate loop identifies the two closest locations s_a^* and s_b^* of the current optimal solution S^*. This step can actually be viewed as the generation of two candidate configurations, each made up of S^* reduced by s_a^* and s_b^*, respectively. Finally, at the inner loop, a simple decision rule is applied, which accepts as the optimal transient configuration one of the two which minimizes the OF:

$$\hat{\phi}_{NPD}(s_i, S_{N-l}\backslash\{s_i\}) = \frac{1}{|S_{N-l}\backslash\{s_i\}|} \sum_{s_j \in S_{N-l}\backslash\{s_i\}} d(s_i, s_j) \tag{2}$$

where $S_{N-l}\backslash\{s_i\}$ is the monitoring network under reduction deprived of s_i and $1 < l < k$ (Ortner et al. 2007). The method stops when the required number of locations (k) has been removed from the original network.

Spatial Simulated Annealing

The SSA is basically structured as a pre-processing stage followed by two nested loops. In the pre-processing, some parameters are estimated, needed to trigger the actual optimization method, namely the initial configuration S_0 and the initial temperature T_0. The outer loop is governed by the temperature and stops when this approaches zero. The inner loop is related to the problem size, that is, if k is defined as the number of locations to be added or removed, the inner loop consists of k iterations for a given temperature value. Within the inner loop, the candidate solutions are generated and subjected to the decision rule (Barca et al. 2015). Concerning the OF, the Average Ordinary Kriging Variance (AKV) has been used. The well-known ordinary kriging variance formulation in a generic unsampled location x_i is (Isaaks and Srivastava 1989):

$$\sigma_R^2(x_i) = \sum_{j=1}^{N} \lambda_j(x_i)\,\gamma(x_j, x_i) - \mu(x_i) \tag{3}$$

where $\lambda_j(x_i)$ are the kriging estimation weights, $\gamma(x_j, x_i)$ is the variogram value for the location pair (x_j, x_i), and $\mu(x_i)$ are the Lagrange multipliers. Consequently, the OF can be written as:

$$\phi_{AKW} = \frac{1}{N}\sum_{i=1}^{N} \sigma_R^2(x_i) \tag{4}$$

It is assumed that a priori knowledge about the spatial law (variogram model) of the variable to be monitored is available. Monitoring-network optimization based on AKV tends to remove locations where the monitoring information is redundant (Barca et al. 2008).

3 Study area

The Alimini lakes are two shallow coastal lakes located in the South-Eastern part of the Apulia Region along the Adriatic Sea coast (Figure 1). Actually, the Northern Lake, named Alimini Grande, is a lagoon, since it is directly connected to the Adriatic Sea through a narrow entrance. The smaller Lake is connected to the other by a natural channel. Both of them are mainly and constantly fed by groundwater recharge, through a number of coastal springs, but also, by surface runoff collected by a network of channels as well as directly by rainfall. In this study, the shallow aquifer has been considered, which is the only one directly connected to the Lakes. From a geological standpoint it is made by Plio-Pleistocene sediments, consisting of an alternating sequence of calcarenites, sands and sandy clays (Margiotta & Negri, 2005). The geometry of aquifer is often hard to determine, since the water lies in limited intervals of permeable rock in a more general context of impermeable deposits. With the aim of investigating the qualitative and quantitative features of the shallow aquifer, 76 agricultural and domestic wells were selected among those located in the neighbourhood of the lakes (Figure 1). This monitoring network is mostly made by dug wells and seldom by drilled wells, whose main characteristics, (e.g. depth, stratigraphy, etc.) are often unknown.

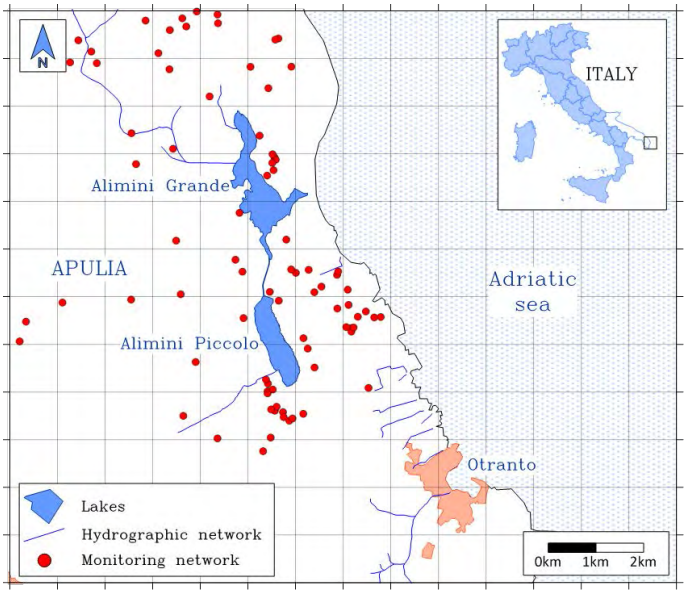


Figure 1: Study area

4 Results and Discussion

The two chosen optimization methods have been applied to the original monitoring network composed by 85 sites. Both the methods have been constrained to halve the network and after the optimization, 41

monitoring sites have been discarded. The network configuration produced by the SSA has an average kriging variance (AKV) which compared with the original network AKV has a percentage of worsening of about 0.9%. As it can be drawn from Figure 2, the two reduced configurations are very similar each other; in fact, they share about the 86% of the same sites. Consequently, it can be expected that, from the representativeness standpoint, the two reduced configurations behave in a similar fashion. In effect, if we try to re-estimate the respective discarded sites values by means of the two reduced network configurations, we obtain the following results:

	MBE	RMSE	RMAE
Greedy Deletion	0.302	1.358	23.161
SSA	0.028	1.367	21.917

Table 1: Summary statistics of network configurations performances.

Analyzing the Table I, we can see that the values estimated by means of SSA show to be significantly less biased than the GD ones. Furthermore, the percentage error (RMAE) is slightly better with respect the GD one. In summary, two out of three indices are very close each other. Consequently, we can conclude that the two reduced configurations perform as expected. A possible explanation of the similar structure and behavior of two configurations can be the extreme clustering of the complete network. Since, the two applied optimization methods tend to intervene on the geometrical configuration of the network; this can explain the obtained result.

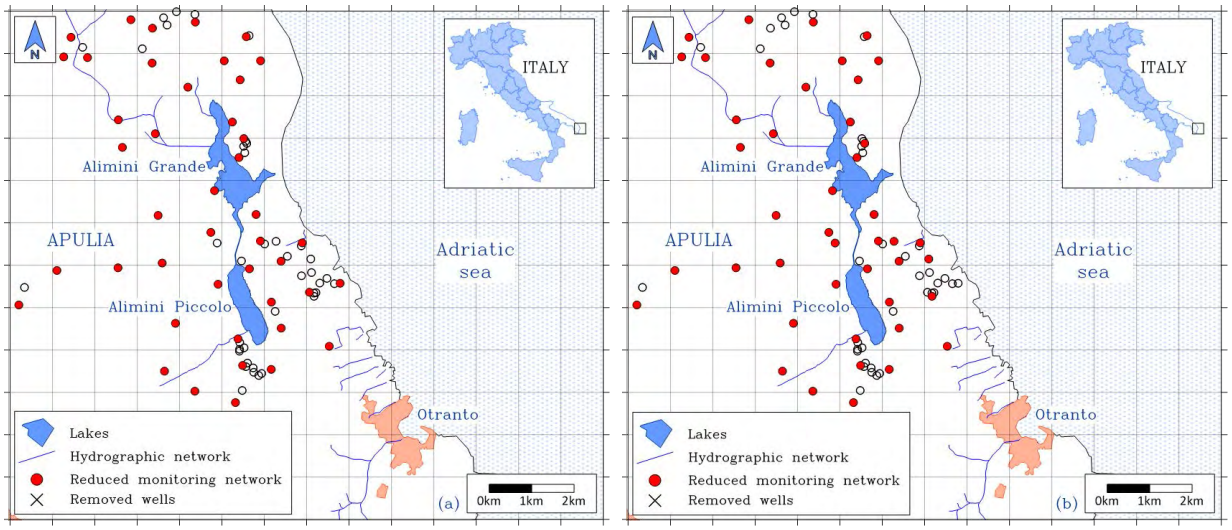


Figure 2: Downsizing results: (a) – Greedy Deletion method; (b) – Spatial Simulated Annealing method.

References

[1] Barca, E., Passarella, G., Uricchio, V., (2008). Optimal extension of the rain gauge monitoring network of the Apulian regional consortium for crop protection. Environ Monit Assess 145(1–3):375–386.

[2] Barca, E., Passarella, G., Vurro, M., & Morea, A. (2015). MSANOS: Data-driven, multi-approach software for optimal redesign of environmental monitoring networks. Water Resources Management, 29(2), 619-644.

[3] Isaaks, E.H., Srivastava, R.M. (1989). An introduction to applied geostatistics. Oxford University Press.

[4] Margiotta S., Negri S. (2005) - Geophysical and stratigraphical research into deep groundwater and intruding seawater in the mediterranean area (the Salento Peninsula, Italy). *Nat. Haz. Earth Sys. Sci.* n.5, pp 127–136.

[5] Nunes, L.M., Cunha, M.C., and Ribeiro, L. (2002). Monitoring network optimization using both spatial and temporal information. In: 3rd Int. Conf. on Decision Making in Urban & Civil Engineering, London.

[6] Wu, Y. (2004). Optimal design of a groundwater monitoring network in Daqing. Env. Geology, 45, 527–535.

Similarity indices of meteo-climatic gauging stations for missing data handling: definition and comparison with the MICE method

E. Barca[1,*], G. Passarella[1]

[1] *Water Research Institute of the National Research Council, Department of Bari, Viale F. De Blasio, 5 70123 Bari, Italy; emanuele.barca@ba.irsa.cnr.it; giuseppe.passarella@ba.irsa.cnr.it*

[*] *Corresponding author*

Abstract. *The meteo-climatic datasets are at the basis of a great deal of studies on environmental state and its consequent management. In this frame, the completeness of meteo-climatic datasets is required for accurate and reliable analysis. Unfortunately, completeness is a rare in practice and, consequently, a preliminary treatment for filling in all gaps is needed. In this work, two intuitive and easy procedures for handling missing data are presented based on the "similarity station" concept. Finally, a comparison between the proposed methods and the Multiple Imputation Chained Equations, which is the state of the art in the field of missing data handling, has been carried out.*

Keywords. *Missing data; Time series; Multiple Imputation Chained Equations; Similarity methods.*

1 Introduction

Cimatic series are rarely complete, usually because of malfunctioning, effects of extreme events on the probes, etc. Consequently, a preliminary formal treatment of the time series is needed in order to fill all the gaps in. Such a treatment is very critical mostly because it is (i) inherently time consuming, particularly for long time series and large amount of missing data; (ii) affected by a high level of uncertainty, particularly for variables irregularly distributed in space and time; (iii) strongly dependent on the missing data mechanism ([2]); (iv) a blind estimation and only a global reliability can be assessed by means of population statistics. At present, a number of robust and powerful methods exist for missing data handling such as the Multiple Imputation Chained Equations (MICE) ([3]) and the Expectation-Maximization (EM) ([1]), which have been designed so that the estimation takes into account the available numerical and distributional information. Such methods revealed their efficacy also in cases where the missing data percentage is particularly severe, overcoming the critical threshold of 15/20%; nevertheless, some authors still claim the need of further investigations to definitively state their reliability ([4]). Furthermore, these methods are practically difficult to be implemented and not very intuitive. In a previous work ([2]) a methodological proposal was presented for a quick and reliable estimation of climatic missing data based on the concept of twin gauging stations. The proposed method is based on the intuitive concept of persistence, in time, of the spatial continuity of the climatic processes. On this basis, a refined and improved methodology is presented for determining similar gauging stations through which estimating missing values. Statistical and topographic properties are combined in order to determine a "similarity matrix". Given a gauging station whose time-series is affected by missing values, these are assessed "combining" the corresponding values of the n most similar stations. The proposed method and MICE were both applied to the rainfall gauging network of the Apulia Region (South-Eastern Italy). Statistical tests on both the estimated time series confirmed a substantial identity between the results of both the methods.

2 Materials and Methods

Usually, matrices of meteo-climatic time-series, measured at different locations of a regional monitoring network, are affected by different rates of missing data. To address this issue, various missing data handling methods have been propose in literature, usually based on the time autocorrelation. Nevertheless, meteo-climatic events, at the considered scale, are notoriously characterized by a strong spatial structure. In practice, we expect similar events in time in "similar stations in space". In this frame, a methodology is proposed in order to assess a degree of similarity of the monitoring network stations. Once a ranked list of "similar stations" has been determined for any considered station, missing values in it can be computed as the average of univariate regression estimations. In this paper two similarity metrics are proposed (equations (1) and (2)): the first approach requires the correlation matrix to be computed. In this case a simple index of similarity $I_R(i,j)$is provided consisting in the determination coefficient (equation (1)). A refinement of $I_R(i,j)$ is also provided ($I_S(i,j)$ in equation (2)) which involves the effective pairs number, $\frac{n_{i,j}}{N}$ and the relative distance and elevation differences, $\hat{d}_{i,j}$and $\widehat{\Delta h}_{i,j}$.

$$I_R(i,j) = R_{i,j}^2 \tag{1}$$

$$I_S(i,j) = \frac{1}{3}\left(\frac{n_{i,j}}{N} \cdot R_{i,j}^2 + \hat{d}_{i,j} + \widehat{\Delta h}_{i,j}\right) \tag{2}$$

$$\hat{d}_{i,j} = \frac{d_{i,j} - d_{k,j}^{max}}{d_{k,j}^{min} - d_{k,j}^{max}} \qquad \begin{matrix} given\ k \\ j=1,\ 2,\ \ldots,\ n \end{matrix} \tag{3}$$

$$\widehat{\Delta h}_{i,j} = \frac{\Delta h_{i,j} - \Delta h_{k,j}^{max}}{\Delta h_{k,j}^{min} - \Delta h_{k,j}^{max}} \qquad \begin{matrix} given\ k \\ j=1,\ 2,\ \ldots,\ n \end{matrix} \tag{4}$$

where i and j represent a pair of monitoring stations; $n_{i,j}$ and N represent the number of pairs shared by i and j and the total length of the time-series, respectively; $\hat{d}_{i,j}$and $\widehat{\Delta h}_{i,j}$ represent the distance and the difference of elevation between i and j standardized with respect to the related variable ranges (equations (3) and (4)). The time–series of the m most similar stations are used for providing m estimations of missing data in i, according to $I_R(i,j)$ and $I_S(i,j)$. Such estimations are finally combined in a single value through the arithmetic average or a weighted average using $I_R(i,j)$ and $I_S(i,j)$ as weights, respectively. An application of the proposed methodologies is presented related to the "Canosa di Puglia" rainfall gauging station time series, which was affected by a severe, fictitious amount of missing data (33%). Estimated values of missing data have been statistically compared with true data and those estimated by the MICE method.

2.1 Study area, monitoring network and rainfall time series

The proposed method has been applied to the monthly total rainfall time series originating from 81 stations irregularly positioned within the Apulia Region (South-Eastern Italy) (Figure 1). In general, rainfall over the Region is characterised by a twofold behaviour depending on the season. Concerning the rainfall regime, it is usually assumed as Mediterranean ([2]). The gauging stations all belong to the meteorological monitoring network of the Regional Hydrographic Services of Land Protection Department. The time series range from January 1931 to December, 2010. The elevation of each station ranges from 2.00 m a.s.l., (Manfredonia station) to 954.00 m a.s.l. (Pescopagano station) and the average distance between the monitoring stations is around 120 km with a standard deviation of 26 km.

3 Results and Discussion

The whole time-series length of Canosa station was made of $N = 960$ monthly total rainfall rates related to the period from January 1931 to December 2010. Values ranging from September 1957 to April 1984 were cut out from the series in order to simulate a long period with missing values (320 values). Table 1 reports the sets of $m = 5$ most correlated (MC) and most similar (MS) stations to Canosa resulting after the indices computation.

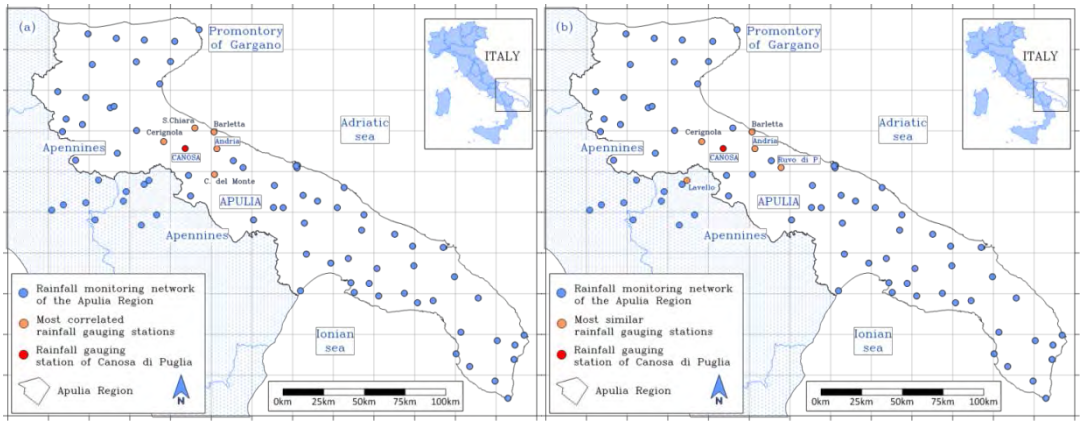


Figure 1: Study area, rainfall monitoring network, Canosa Station and most similar stations according to $I_R(i,j)$in (a) and $I_S(i,j)$ in (b).

Rain gauge Station	**MC/MS**	$R^2_{i,j}$	$d_{i,j}$	$\Delta h_{i,j}$	$n_{i,j}$	$I_S(i,j)$
Masseria Santa	MC	0.789	13.9	145.0	328	-
Cerignola	MC/MS	0.782	13.9	30.0	634	0.852
Andria	MC/MS	0.742	19.5	3.0	634	0.845
Barletta	MC/MS	0.731	20.4	124.0	597	0.793
Castel del Monte	MC	0.718	24.0	371.0	509	-
Lavello	MS	0.706	29.6	159.0	625	0.774
Ruvo di Puglia	MS	0.676	37.4	106.0	621	0.774

Table 1: Summary of parameters involved in the indices computation.

The third and the last columns of Table 1 report the correspondent values of $I_R(i,j)$ and $I_S(i,j)$. Monthly missing values of Canosa have been estimated five times by means of linear univariate regression using each of the most correlated station. Finally, a single value has been computed by averaging the five estimated values. Figure 2 shows the plots of estimated versus true observed values, cut out previously. In particular plot a) refers to the results of the most correlated stations approach, plot b) to that of the most similar approach and finally, plot c) shows the results obtained estimating the missing values by means of the well-known MICE method. Figure 2 is not decisive in order to establish what of the three methods prevails over the others. In fact, the goodness of fit coefficient shows a slightly better value for the "most correlated" approach than the others, while the coefficient and the shape of the regression line seem to indicate better results from MICE. In any case the differences seem to be negligible. Even the error statistics, reported in Table 2, do not indicate the best approach, clearly. The values of the Mean Bias Error (MBE), of the Root Mean Squared Error (RMSE) and of the Relative Mean Absolute Error (RMAE) in Table 2 are very close each other.

4 Conclusions

Four methods have been proposed to estimate missing values in long time series of rainfall measures. The methods substantially propose a univariate regressive estimation of the missing values, in a given rainfall gauging station of a regional monitoring network, based on a set of "similar" stations located in the neighboring. Two indices of similarity have been proposed: the coefficient of determination

($I_R(i,j)$) between the dependent station i and the other stations j and the index $I_S(i,j)$ computed combining some statistical and topographic properties of the pairs (i,j). Both the proposed indices range from 0 to 1.

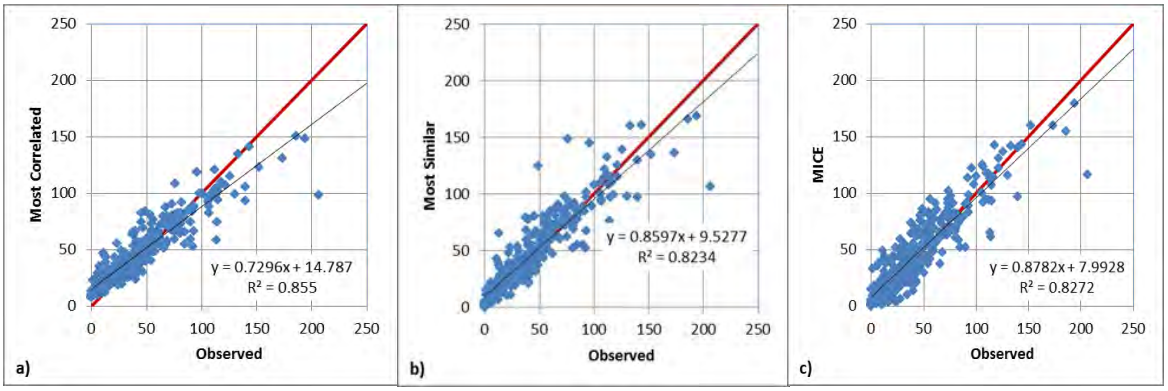


Figure 2: Observed vs. Estimation: a) Most Correlated; b) Most Similar; c) MICE method.

Once a ranked list of correlated/similar stations has been determined, missing values in i are computed as an average of univariate regression estimations. A case study related to the "Canosa di Puglia" gauging station, locate in Apulia Region (South-Eastern Italy), has been presented. In Canosa, monthly total rainfall rates have been measured continuously from 1931 to 2010. A set of 320 values (i.e. 33% of the whole dataset) has been cut out from the series and estimated with the proposed methods. Finally, provided that the MICE method is one of the most reliable for data missing estimation, available in literature, it has been used as benchmark to assess the performances of the proposed methods.

Method	**MBE**	**RMSE**	**RMAE**
Most Correlated	2.54	15.00	0.57
Weigthed Most Correlated	2.54	14.96	0.57
Most Similar	3.21	15.80	0.43
Weigthed Most Similar	3.17	15.72	0.43
MICE	2.47	15.50	0.54

Table 2: Summary statistics of models estimation error.

The results demonstrate a clear equivalence of all the methods, in terms of estimation error statistics and goodness of fit. In conclusion, considering meteo-climatic time-series missing data issue, the proposed methods seem to behave similarly to the most celebrated MICE; however, the procedural straightforwardness of the proposed methods can lead to prefer them instead of other methods well-known for their effectiveness but, undoubtedly, more complex in terms of computational efforts. Monthly total rainfall and monthly mean temperature time-series related to the 81 gauging stations of the regional meteo-climatic monitoring network have been filled in using the Weighted Most Correlated approach and the aforementioned aridity indices have been computed, spatialized and mapped for managerial uses.

References

[1] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. Series B, **39**(1), 1–38.
[2] Lo Presti, R., Barca, E., Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental monitoring and assessment*, **160**(1-4), 1-22.
[3] Rubin, D.B. (1987*). Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.
[4] Schafer, J.L., Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological methods*, **7**(2), 147.

Statistical analysis of acoustic data. Combining objective and subjective measures.

C. Bartalucci, F. Borchi, M. Carfagni[1], M.S. Salvini and A. Petrucci[2,*]

[1] *Department of Industrial Engineering, University of Florence, via di S. Marta 3, 50139 Firenze, Italy; chiara.bartalucci@unifi.it , francesco.borchi@unifi.it, monica.carfagni@unifi.it*
[2] *Department of Statistics, Computer Sciences, Applications "G. Parenti", University of Florence, Viale Morgagni 59, 50134 Firenze, Italy mariasilvana.salvini@unifi.it, alessandra.petrucci@unifi.it*
[*] *Corresponding author*

Abstract. *This paper presents a statistical approach to analyze a set of data collected by QUADMAP (QUiet Areas Definition and Management in Action Plans), a LIFE+2010 Project on Quiet Urban Areas (QUAs). This Project aims to define a method regarding identification, outlining, characterization, improvement and managing of QUAs as meant in the Environmental Noise Directive 2002/49/EC. The project will also help to understand the definition of a QUA, the meaning and the "added value" for the city and their citizens in terms of health, social safety and lowering stress levels. At the beginning of 2013 the first version of a methodology to select, analyze and manage QUAs has been produced and subsequently applied in ten pilot areas chosen in Firenze, Bilbao and Rotterdam. During the analysis phase, quantitative (noise maps and acoustic measurements) and qualitative (end-users questionnaires, general and non-acoustic information) data have been collected and examined. Once the ante-operam phase of analysis has been completed, the interventions' realization in the pilot areas started and was followed by post-operam surveys. Logistic models are applied to part of the quantitative (noise measurements) and qualitative (results of interviews about noise perception in an urban setting) data collected in the city of Florence during the ante-operam phase. The results underline the importance of the survey's design, in order both to obtain information combining the different type of data and to be able to evaluate the net effect of single variables.*

Keywords. *Logistic models, Acoustic data, Objective measures, Subjective measures*

1 Introduction

The European Directive 2002/49/EC on the Assessment and Management of Environmental Noise (END) was adopted to define a common approach to avoid, prevent or reduce the harmful effects due to noise exposure and to preserve the environmental noise quality where it is good.
QUADMAP Project started in September 2011 with the final aim of developing a complete, practical and demonstrated methodology to select, analyse and manage QUAs (Quiet Urban Areas). At the beginning of 2013 a first version of the methodology was drafted and, as a consequence, it started to be tested in the pilot cases identified by the Project partners [1]. To this aim the following typologies of case studies have been chosen: six schoolyards in Florence, one square and a peri-urban green corridor in Bilbao and two public parks in Rotterdam. The application of the methodology to the pilot cases during the ante-operam

phase led to its updating, while the final optimization has been achieved after the interventions' realization and the post-operam phase have been carried out at the beginning of 2015 [2]. Consequently, guidelines to facilitate the application of the methodology have been developed [6]. QUADMAP Project promoted the application of a participatory approach, in terms of questionnaires submitted to end-users, aimed at integrating the people's perception with objective acoustic measurements.

2 Data

By applying tools described in the methodology [2, 6], the following typologies of data have been collected by QUADMAP in each pilot area:

- quantitative data (noise maps, short and long term measurements, wave recordings);
- qualitative data (end-users questionnaires, general and non-acoustic information).

For the purpose of this study, short term measurements and end-users questionnaires collected in the six pilot cases located in Florence during the ante-operam phase are deeper evaluated by using statistical analysis and models with the main aim to understand the effect of each considered parameter.

The format of the end-users questionnaire has been developed by the project partners, it is very broad and includes more than one hundred questions, structured with open and closed questions regarding both the acoustical and general perception of QUAs [2, 6].

Short term measurements are based on the Time History of sound pressure levels. They are carried out at the same time as each end-users questionnaire and have a duration of 15-30 minutes. Collected data have been considered in a comprehensive manner, without taking into account the division by school and age of the respondent, in order to have an overview of the selected cases.

Among available statistical models, logistic regression and ordinal logistic models have been evaluated.

3 Methods

After collecting data and before performing synthetic models, firstly a descriptive analysis of the sample has been carried out. According to this analysis, the structure of the examined group results to be not completely consistent with the overall population. The majority of respondents is female (61.6%) and the age distribution is highly skewed; in fact, over 60% are under 20 years and about 35% have more than 30. Students, children and adolescents under 20 years old who attend school gardens form the sample size for almost two-thirds, while the remaining part is divided into various trades. This particular composition, dominated by children, can have an impact on the responses to the questionnaire which are in fact more suitable to older age than for young people and in some cases may have been misled. Therefore, the structure of the sample may also influence the results of the statistical models that have been applied to the data set. Independent variables considered for the models are both quantitative and qualitative, while the only considered dependent variable is a qualitative one, i.e. the addressed question is "I value this area in general as good", with possible answers from 1 to 5 (Fully disagree, Disagree, Neutral, Agree, Fully agree). Concerning qualitative independent variables, they have been selected among the questions of the end-user questionnaire and are "Referring to this area, I perceive each of the following items as pleasant: Air quality, Safety, Well-maintenance, Services and equipment (benches, playing areas ..), Accessibility, Acoustic environment, Natural elements (green areas, water, birds, etc.), Climate (humidity, brightness, wind), Visual aspects, Smells". The answer to these questions is to express an opinion on the different aspects of the quality of the area, translated into a score, on a scale from 1 to 5 (not pleasant at all, quite unpleasant, pleasant enough, pleasant, very pleasant). Information acquired from the noise measurements constitute the objective part of the database. Based on short term measurements, parameters such as LAeq (equivalent continuous sound pressure level), L10-L90 (sound level exceeded respectively for 10% and 90% of the measurement time) and LA50 (sound level exceeded for 50% of the measurement time) are evaluated and taken into account as independent variables for the statistical models.

4 Results

Firstly, a simple logistic regression model has been applied, using as dependent binary variable the general perception of area and as quantitative explanatory factor only the LA50. Results suggest that, among those evaluated from short term measurements, the most appropriate parameter to describe the perception of users is the LA50 [2]. Afterwards, more complex logistic regression and ordinal regression models have been implemented, considering as dependent factor still the general perception of the area and as independent variables all the qualitative ones above cited and, according to results obtained with the

simpler models, the LA50 as unique quantitative factor. Before running the logistic model, according to its structure, answers from 1 to 5 (Fully disagree, Disagree, Neutral, Agree, Fully agree) to the question chosen as dependent variable have been reclassified in two modalities: agreement (correspondent to: Neutral, Agree, Fully agree), disagreement (correspondent to: Fully disagree, Disagree). In table 1 outputs of this last model are reported, stressing that only a few variables show statistically significant influence on the general perception of the area. In particular, variables evaluated as significant are all qualitative and show signs of the coefficients in the expected sense (e.g. the air quality tends to increase the positive perception of the area when evaluated as pleasant or very pleasant), while the LA50 appears to be not statistically significant when considered together with the qualitative variables.
Much more explicit are the results of the ordinal logistic model, where the dependent variable is left in five modalities (see table 2).

Parameter	Modalities	Degree of Freedom (DF)	β	Standard Error (SE)	Wald Chi-Square	Pr > $\chi 2$
Air quality	3	1	0.7373	0.4387	2.8246	0.0928
Air quality	4	1	1.1032	0.6462	2.9146	0.0878
Maintenance	5	1	1.5989	0.8694	3.3820	0.0659
Climate	2	1	-1.3714	0.6131	5.0029	0.0253
LA50		1	0.0101	0.0456	0.0490	0.8249

Table 1. Logistic regression models: parameters estimation

Parameter	Modalities	Degree of Freedom (DF)	β	Standard Error (SE)	Wald Chi-Square	Pr > $\chi 2$
Intercept**	5 very pleasant	1	-3.8362	1.0763	12.7047	0.0004
Intercept	4 pleasant	1	-1.8951	1.0619	3.1847	0.0743
Intercept	3 quite pleasant	1	0.1074	1.0591	0.0103	0.9192
Intercept	2 not so pleasant	1	2.5527	1.1026	5.3596	0.0206
Smells	s*	1	0.2649	0.1427	3.4452	0.0634
Visual aspects	s*	1	0.2183	0.1346	2.6312	0.1048
Climate	s*	1	0.2520	0.1868	1.8197	0.1773
Natural elements	s*	1	0.4896	0.1676	8.5367	0.0035
Acoustic environment	s*	1	0.4252	0.1369	9.6395	0.0019
Availability	s*	1	-0.1009	0.1685	0.3584	0.5494
Services	s*	1	0.3412	0.1158	8.6834	0.0032
Maintenance	s*	1	0.3812	0.1354	7.9246	0.0049
Security	s*	1	0.3179	0.1614	3.8782	0.0489
Air quality	s*	1	0.4488	0.1581	8.0599	0.0045
LA50	s*	1	0.00907	0.0178	0.2589	0.6109

Table 2. Ordinal logistic models: parameters estimation

In table 3 the goodness of fit of both models to the database is evaluated.

Model	N° of significant variables	Pseudo R-square	H&L GOF (p-value)
Logistic (ordinal)	7	0.3947	
Logistic	4	0.3545	0.6062

Table 3. Goodness of fit of the models

5 Conclusions

In this paper descriptive analysis and statistical models have been evaluated, starting from qualitative and quantitative data achieved during the ante-operam phase in the pilot cases selected in Florence to analyze noise objective and subjective variables.

According to both descriptive analysis and statistical models, the users of the six school gardens in Florence discriminate the area mainly on the basis of the perception of the air quality and of the well-maintenance. The role of the quantitative variables is found to be, however, quite marginal. Also in the literature [3, 4, 5] examples are present of how quantitative information, although they are the most obvious to detect, can be misleading if they are the only considered ones. In particular, it can be seen that the measured noise levels can hardly be associated to the users' perception of the external area. In fact, it may happen that the sound environment is negatively judged in case of low noise levels and vice versa. This is probably explained by the fact that, often, even if they perceive high levels of noise, they do not generate annoyance because they are unconsciously perceived as an integral part of the garden itself. In addition, since the sample was prevalently composed by children sometimes also under six years old, it has been quite difficult to ask them especially about the perception of the acoustic environment. Consequently, during the post-operam surveys a simplified version of the questionnaire has been developed and submitted in schoolyards.

About the comparison between the two typologies of tested models, it can be concluded that the ordinal logistics models show a higher number of significant coefficients if compared to logistic regression ones and, in addition, in the first case many variables can be interpreted as determinants of the general perception of the area as good.

References

[1] Bartalucci, C., Bellomini, R., Borchi, F., Carfagni, M., Governi, L., Luzzi, S. and Natale, R. (2013). LIFE+2010 QUADMAP project (Quiet Areas Definition and Management in Action Plans): the proposed methodology and its application in the pilot cases of Firenze, Proceedings of the 42st International Congress on noise, Innsbruck, Austria, 15-18 September, 2013.

[2] Carfagni, M., Bartalucci, C., F. Borchi, L. Governi, Petrucci, A., Weber, M., Aspuru, I., Bellomini, R., Gaudibert, P. (2014). LIFE+2010 QUADMAP Project (QUiet Areas Definition and Management in Action Plans): the new methodology obtained after applying the optimization procedures, Proceedings of the 21st International Congress on Sound and Vibration, Beijing, China, 13-17 July, 2014.

[3] Curcuruto, S., Asdrubali, F., Brambilla, G., Silvaggio, R., D'Alessandro, F., Gallo, V. (2011). Socio acoustic survey and soundscape analysis in urban parks in Rome, Proceedings of the 11th ICBEN Congress, London, United Kindom, 24-28 July, 2011.

[4] Carvalho APO, Morgado AEJ, Henrique L. Relationship between subjective and objective acoustical measures in churches. (1997). *Build Acoust* **4**, 1–20.

[5] Engel, M.S., Hochsteiner De Vasconcelos Segundo E., Trombetta Zannin P. H. (2014). Statistical analysis of a combination of objective and subjective environmental noise data using factor analysis and multinomial logistic regression, *Stoch Environ Res Risk Assess* **28**, 393–399.

[6] QUADMAP Project, (2015). Guidelines for the identification, selection, analysis and manage-ment of quiet urban areas: http://www.quadmap.eu

Development of biogas and management of the nitrates in Veneto

P. Belcaro[1,*] and F. Schenato[2]

[1] *pierantonio.belcaro@regione.veneto.it*
[2] *federica.schenato@yahoo.it*
**Corresponding author*

Abstract. *This study aims to analyze the development of biogas from an agricultural and animal husbandry matrix in the Veneto in order to evaluate its effects in terms of the involvement of animal biomass and production of nitrates. The analysis showed that the use of the digestate, from the process of biogas production, such as fertilizer, reduces the final quantity of nitrogen distributed directly on the ground in comparison with that derived from the traditional use of livestock and agricultural effluents, thereby reducing the possibility of groundwater pollution. The data on the location of the plants and the production of biogas, in connection with those related to the management of livestock manure and the production of nitrates, form a base of information that is useful for designing policies aimed at protecting the areas vulnerable to nitrates in the Veneto region.*

Keywords. *Biogas; Nitrates; Nitrate Vulnerable Zones; Groundwater.*

1 The development of biogas in Veneto[1]

Energy policies implemented by European countries are aimed at diversifying energy supply sources, increasing energy efficiency and developing renewable energy. The Directive 2009/28/EC of the European Union set a target of 20% of energy production from renewable sources in total gross final consumption to be achieved by 2020 in the EU and a target of 17% for Italy. With various acts of programming, such as the National Action Plan for Renewable Energy in Italy (PAN) of 30 June 2010, the DM 15 March 2012 of the Minister of Economic Development, the so-called Decree Burden Sharing, for the splitting of the Italian target between different regions, and different regional energy plans, the possible development of each renewable source was defined in accordance with the overall targets set at the top level. In this sense, a significant role is played by bio-energy and particularly by biogas, which saw a rapid development in recent years throughout Italy and the Veneto thanks to the strong incentives that policy put in place. We have moved, at the national level, from a biogas production of 1,336.3 GWh in 2006 to a production of 4,619.9 GWh in 2012. Among the various matrices used, the products from agriculture and forestry carry a significant weight, contributing 54.8% (2012 data) of the total production compared with 8.3% in 2006. The same increasing change applies to the animal wastes related to animal husbandry with an 11.2% (2012 data) contribution; a significant increase when compared with 3.3% of 2006. In 2012 in Veneto, with 197 plants, 571.8 GWh or 12.4% of national production were produced from biogas, with a very diversified supply from various provinces.

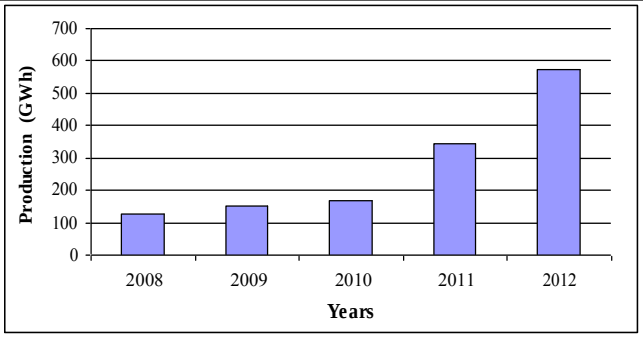


Figure 1: Production of energy from biogas in the Veneto (Source: GSE)

Provinces	Production from RES	Production from bio-energy	Production from biogas
Belluno	2.193.4	212.6	4.5
Padova	543.5	250.1	169.3
Rovigo	452.9	87.5	86.9
Treviso	1,0029	37.5	32.0
Venezia	452.9	325.1	109.8
Verona	1,184.0	136.4	132.7
Vicenza	640.5	875	36.6
Total Veneto	6,470.2	1,136.7	571.8

Table 1: Production of energy from Renewable Energy Sources (RES), bio-energy and biogas in GWh in the Veneto provinces in 2012 (Source: GSE)

As with all energy sources, there is the problem with assessing the impact of biogas on the environment[1] and health. There were no major objections expressed so far compared to plants licensed or in operation, but proper verification and reporting on pollution associated with this type of energy production is useful in enabling the public administrators to take decisions advisedly and to prevent onset of so-called NIMBY syndrome.

1 This paragraph was edited by Federica Schenato

2 The management of nitrates connected at biogas plants[2]

Fertilization performed with the distribution of livestock effluents or with the use of digestate, i.e. the product resulting from the anaerobic digestion implemented in the process of biogas production, causes the release of nitrogen in the soil, which, if not absorbed by plants, can represent a potential source of pollution for groundwater and surface water. In order to protect water resources, the European Union enacted Directive 91/676/EEC, the so-called Nitrates Directive, which established the basic principles for the regulation of these activities in order to promote the proper use of agricultural practices. This Directive, which provides for the identification of Nitrate Vulnerable Zones (NVZ) and the implementation of Action Programs, is implemented in Italy by the Legislative Decree n. 152/1999, later replaced by Legislative Decree n. 152/2006 (Environmental Code). Subsequently, the Regional Government of Veneto with D.G.R. n. 2495/2006 approved the first Action Program for vulnerable zones by nitrates from agricultural sources, by regulating the activities of spreading of livestock manure for vulnerable zones and for the remaining agricultural areas; it was followed by the second Action Program covering the period 1st January 2012 - 31st December 2015, approved by D.G.R. n. 1150/2011. This Program is also effective in the territories classified as vulnerable zones in the drainage basin of the Venice lagoon, the Province of Rovigo, the municipality of Cavarzere, the hundred municipalities of the high plain, the municipalities of Lessinia and reliefs in the right Adige. These large areas of land correspond to about 60% of the surface of the entire region. It should be remembered that the current

[1] An interesting model for the analysis of the environmental impact of biogas plants, limited to atmospheric emissions, is contained in the analysis conducted by TIS - Techno Innovation Alto Adige S.C.p.A..

legislation provides different limits for the maximum amount of nitrogen by animals that can be used in spreading the field: 170 kg/ha per year in NVZ and 340 kg/ha per year in Ordinary Zone (OZ).

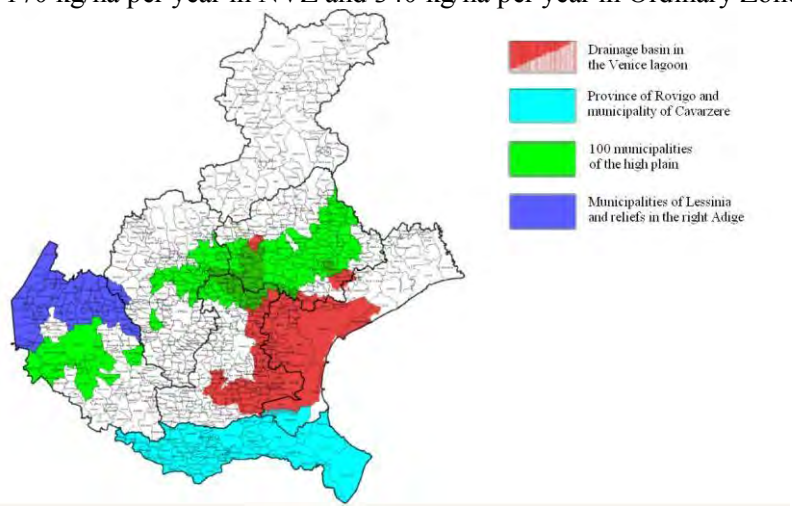


Figure 2: Map of NVZ (Source: Regione Veneto)

Let us now consider the data related to the nitrogen input and output from biogas production plants fed in whole or in part from animal manure, which results from the nitrates database built according to spreading plans and agronomic use manure communications. These communications are for installations with power up to 1 MW electric and 3 MW thermal.

	Zone type	NVZ	OZ	Total
	No. Plants	61	58	119
Nitrogen in input in the plants	Nitrogen from zootechnic manure (kg/yr)	2,301,419	1,844,548	4,145,967
	Nitrogen from other biomasses (kg/yr)	1,708,597	2,510,105	4,218,702
	Total nitrose (kg/yr)	4,010,016	4,354,653	8,364,669
Nitrogen in output in the plants	Nitrogen from zootechnic manure (kg/yr)	1,816,079	1,701,819	3,517,898
	Nitrogen from other biomasses (kg/yr)	1,419,464	2,312,929	3,732,393
	Total nitrogen (kg/yr)	3,235,543	4,014,748	7,250,291
	% abatement of nitrogen from zootechnic manure	-21.1%	-7.7%	-15.1%
	% abatement of nitrogen from other biomasses	-16.9%	-7.9%	-11.5%
	% abatement of total nitrogen	-19.3%	-7.8%	-13.3%

Table 2: Nitrogen treated by the plants for biogas production in 2013. (Source: Based on data from the Regione Veneto - Sezione Agroambiente)

The data show a general decrease in the amount of nitrogen, which, although not caused from the anaerobic digestion process, but rather from the accessories treatments which underwent the digestate in output, goes in the direction of the reduction of the amount of nitrogen used directly in the fertilization of soils, compared to the amount that would have been available without the presence of digesters. This decrease was more pronounced in NVZ compared to OZ, and for nitrogen by animals, compared to that produced by other biomasses. In confirmation of these evaluations, it should be noted that in the 16 plants that use only zootechnic manure the decrease of the amount of nitrogen reaches 18.3%.

While it is not possible to establish a strict correlation between the development of biogas production from the treatment of livestock manure and the trend of the content of nitrates in water in Veneto, we can still make some indicative valuations.

In fact, in accordance with the Strategic Environmental Assessment and the requirements of the second Program of Action, to check the progress of nitrate concentrations and the environmental state of water bodies, the Sezione Agroambiente of the Region prepares special annual reports of environmental monitoring. The "Monitoring Report 2013 VAS nitrates" shows the data on average concentrations for the year 2012 and 2013 of nitrate in groundwater by sampling stations in the NVZ and OZ. These data reveal that the average concentration was kept constant at 16 mg/l in the NVZ and decreased from 8.9 mg/l in 2012 to 8.2 mg/l of 2013 in OZ. The Report shows that in the region, 83% of groundwater records in 2013 the chemical status of "good" and 17% "poor" rating due to the 3% for nitrates and to 14% for other pollutants. The same parameters referred to the NVZ make record 80% of "good" and 20% "poor", the latter due for 5% to nitrates and for 15% to other pollutants.

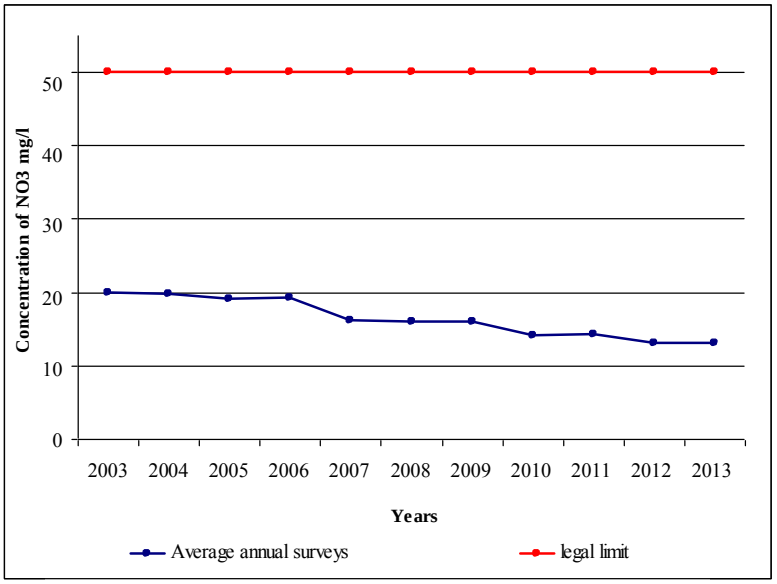


Figure 3: Average concentrations measured at the sampling points of groundwater (Source: Based on data Arpav)

The analysis presented in this work, albeit with the necessary caution due to the fact that the models have not yet completely defined the relationship between nitrogen and natural resources, it can be concluded that the presence of plants for biogas production leads to a reduction of over 10% of the amount of nitrogen released into the ground and that, in recent years, there has been a reduction in the average value of the concentration of NO_3 in sampling points of groundwater.

2 This paragraph was edited by Pierantonio Belcaro

References

[1] GSE (2014). *Rapporto Statistico Impianti a fonti rinnovabili - anno 2012*. Roma.

[2] Regione del Veneto - Sezione Agroambiente (2014). *Report di Monitoraggio VAS nitrati - anno 2013*. Regione Veneto. Venezia.

[3] TIS - Techno Innovation Alto Adige S.C.p.A. (2011). *Analisi energetica, ambientale ed economica di impianti a biogas in Provincia di Bolzano - Relazione conclusiva*. Provincia Autonoma di Bolzano. Bolzano.

Fauna characterization of a cold-water coral community network along the Apulian coasts by Bayesian mixed models

C. Calculli[1,*], G. D'Onghia[2], N. Ribecco [3], P. Maiorano[2], L. Sion[2], A. Tursi[2]

[1] *CoNISMa Local Research Unit Bari - Italy, calculli.enza@gmail.com*
[2] *Department of Biology, University of Bari Aldo Moro, via E. Orabona 4, 70125 - Italy, gianfranco.donghia@uniba.it, porzia.maiorano@uniba.it, letizia.sion@uniba.it, angelo.tursi@uniba.it*
[3] *Department of Economics and Mathematical Methods, University of Bari Aldo Moro, Largo Abbazia Santa Scolastica 53, 70124 Bari, Italy, nunziata.ribecco@uniba.it*
[*] *Corresponding author*

Abstract. *The exploration of a cold-water coral (CWC) community network connecting the Southern Adriatic fish populations with those of the Northern Ionian Sea has many challenging implications involving biodiversity conservation and fisheries management. To characterize the species assemblages of the CWC community network, we analyze the size of six mostly abundant species to highlight the main differences among five CWC areas along the Apulian coasts. Data are surveyed by experimental longlines casted in all CWC areas between 2013 and 2014. Bayesian Generalized Additive Mixed Models (GAMMs) are applied to analyze the variation of the fishes length in the five CWC areas according to species and covariates such as depth and abundance. GAMMs allow to account for various effects of variability components on the overall length of fishes. Appropriate smooth functions are available to describe the physical effect of each covariate and specific random area effects are allowed to induce correlation among individuals of all species captured in the same area. Parameter estimation is carried out by maximization of the joint posterior probability distribution, where marginal posterior probabilities associated to specific model terms are obtained by a* spike-and-slab *prior structure, that can be viewed as a scale mixture of Gaussians. This approach is implemented in the R package spikeSlabGAM, able to deal with most common distributional assumptions and allowing efficient variable selection and model choice. Results show that the size of fishes is collectively affected by random effects of the CWC areas and by smooth effects of their depth and abundance.*

Keywords. *Bayesian GAMMs; Random effects; Cold-water coral community network; Fish length*

1 Introduction

The Apulian continental margin (Central Mediterranean) is characterized by the presence of a network of cold-water coral (CWC) communities connecting the Adriatic with the Northern Ionian Sea fishes populations [1]. The water masses flowing from the Adriatic into the Ionian Sea allow to connect areas in which cold-water corals can thrive [5]. Understanding the functioning of this CWC community network is a crucial issue because of its importance, both in terms of the habitat conservation and for

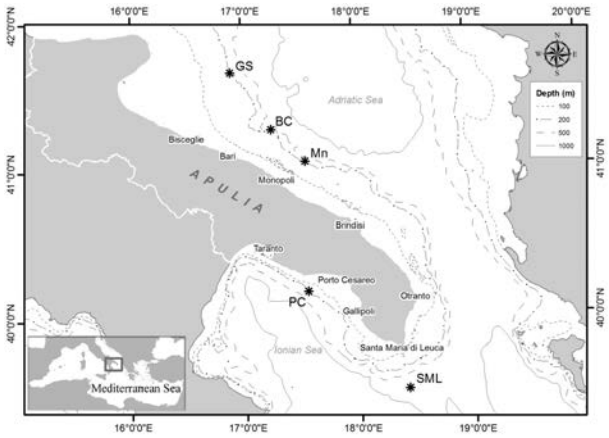


Figure 1: CWC community network along the Apulian coasts

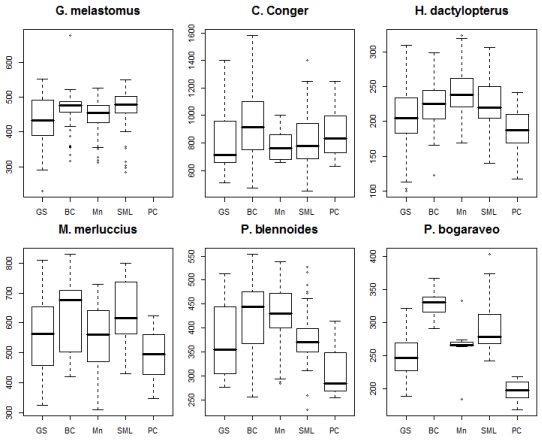


Figure 2: Size species distributions for five CWC areas

the management of the living resources impacted by commercial fishing [2, 3]. As a consequence, the protection and conservation of CWC areas have become a recognised priority whereby effective management and biodiversity objectives measures might be developed. The aim of this paper is to characterize the species assemblages of the CWC community network existing in the Southern Adriatic-Northern Ionian. We consider the fishes length as an indicator for investigating the main differences among the community network areas. An approach based on Generalized Additive Mixed Models (GAMMs) is proposed. These represent a wide class of models in the context of structured additive regression [4] and can be considered an additive extension of Generalized Linear Mixed Models using additive non parametric functions to model covariate effects and accounting for overdispersion and correlation by random terms. GAMMs are suitable for nested and crossed designs and are applicable to clustered, hierarchical and spatial data [8]. For this reason, GAMMs are particularly suited for analyzing the case study data grouped by areas. The estimation of parameters associated to specific model terms is obtained in a Bayesian perspective, following the approach suggested in [9] which is based on a flexible priors structure, named *spike-and-slab*, and implemented in the related **spikeSlabGAM** `R` package for fully Bayesian variable selection and model choice.

2 Material and Methods

The CWC areas investigated are distributed along the Apulian coasts as shown in Figure 1. A total amount of five areas are located between the Southern Adriatic (Gondola Slide - GS, Bari Canyon - BC, off Monopoli - Mn) and the Northern Ionian sea (the CWC province Santa Maria di Leuca - SML, off Porto Cesareo - PC). We consider data from experimental longlines surveys collected among 2013 and 2014, minimizing the impacts on the sea-floor and benthic fauna. In particular, to study the species assemblages, the sizes of six mostly abundant fish species in each area are considered: *Galeus melastomus, Conger conger, Helicolenus dactylopterus, Merluccius merluccius, Pagellus bogaraveo, Phycis blennoides*. Boxplots in Figure 2 report the distributions of the lengths for each species and area, showing for some species (e.g. *P. bogaraveo* or *P. blennoides*) differences in the sizes between areas. To account for various effects of variability components on the overall length of fishes in the five CWC areas, we propose a GAMM approach modelling dependence of the size response on the fish species, depth and abundance covariates. Generally, given a set of covariates $x_j (j = 1, \ldots, p)$, the distribution

	$P(\gamma=1)$	*pi*	
sm(abund)	0.467	-0.005	*
lin(depth)	0.999	0.069	***
fct(species)	1.000	0.940	***
rnd(area)	0.274	-0.004	*

Table 1: Marginal posterior inclusion probabilities and terms of importance *pi* for each model component. *: $P(\gamma=1)>0.25$, **: $P(\gamma=1)>0.5$, ***:$P(\gamma=1)>0.9$.

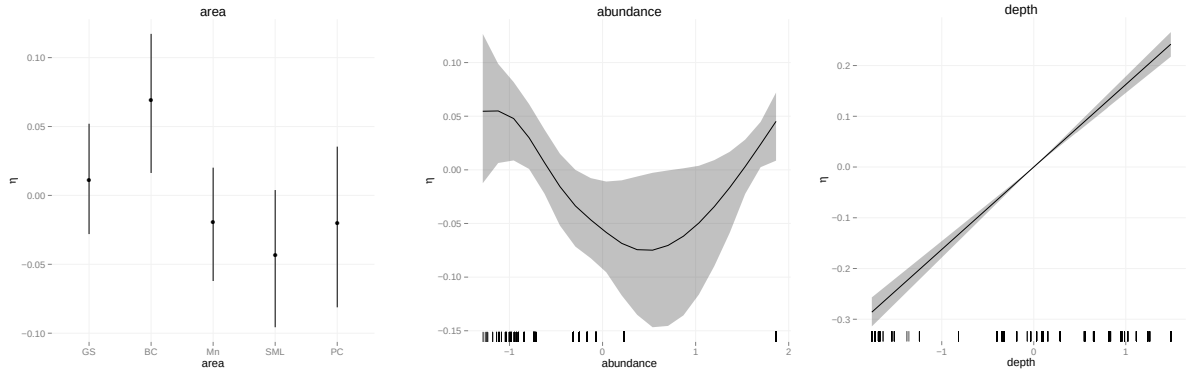
of the response y in a GAMM belongs to the exponential family and the conditional expected value $E(y|x_1,\ldots,x_p)=h(\eta)$ is determined by the additive predictor η including a wide variety of model terms as linear terms, nominal or ordinal covariates, smooth functions of continuous covariates (spline, tensor product or varying coefficient term), random effects (subject-specific intercepts and slope) and interactions between the different terms. For the case study, let y_{it} denote the standardized length of the t-th individual observed in the i-th area, with $t=1,\ldots,1605$ and $i=1,\ldots,5$. Here we assume that the standardized response has a Gaussian distribution and that the additive predictor is related to the expected value of the data through the identity link function (h):

$$\eta_{it}=\beta_1 f_1(abund_{it})+\beta_2 f_2(depth_i)+\beta_3 species_t+\beta_4 area_i \tag{1}$$

where functions $f_{1,2}$ are unknown smooth functions respectively associated to the abundance and depth covariates, while the covariate species is a six levels factor forced to be a fixed component. Coefficients $\beta_j, j=1\ldots,4$ related to each model term are to be estimated and the area is introduced as a random effect. As a consequence the model intercept is allowed to vary for each CWC area accounting for correlation between observations belonging to the same area. In a Bayesian perspective, inference on GAMM parameters can be carried out considering all model components as random variables supplemented by appropriate prior assumptions. The estimation of marginal posterior probabilities of terms in Eq. (1) can be achieved using *spike-and-slab* priors (peNMIG) which are bimodal priors for the regression coefficients that are decomposed in a two component mixture of a narrow spike around zero and a slab with wide support for the marginal prior of the coefficients themselves [7]. The generic peNMIG prior for β_j can be viewed as a scale mixture of Gaussians, while the variance prior common to all β coefficients is Inverse Gamma. The posterior mixture weight for the spike component of a specific coefficient can be interpreted as the posterior probability of its exclusion from the model. The proposed approach is implemented by means of **SpikeSlabGAM** `R` package which is suitable for Gaussian, Binomial and Poisson responses and features efficient model selection as well as model choice.

3 Results

Table 1 shows the MCMC results of 3 chains each running 10000 iterations after a burn-in phase of 500 and thinning by 5. The graphical inspection of trace plots highlights strong evidence of good mixing and convergence for the 3 chains and all parameters involved with the proposed model, thus reaching the target posterior distribution. The shape of each effect is chosen according to the nature of the corresponding covariate, thus the abundance is associated with a univariate penalized spline (B-spline), the depth with a linear trend, the species with a factor effect and the area with a random intercept. The posterior probabilities of inclusion for this terms are all significant and the shape of fitted effects are reported in Figure 3.



(a) (b) (c)

Figure 3: Posterior means and pointwise 80% credible intervals for some fitted model components.

The abundance has a non linear decreasing effect on the length of fishes (smaller individuals are more aggregated), while the depth has a linear increasing trend according to the bigger-deeper phenomenon. The fitted random effect area shows as the lengths differ in the five CWC areas as a consequence of bathymetric ranges investigated. The species factor has a natural expected effect on the length of all fishes (not reported in Figure 3).

In this paper an efficient alternative to deal with GAMMs is proposed. The approach based on the Bayesian estimation of model parameters is very appealing allowing to specified complex models concisely. The handling of many kind of effects makes this method particulary recommended for the analysis of nested or hierarchical structured ecological data.

References

[1] Angeletti, L., Taviani, M., Canese, S., Foglini, F., Mastrototaro, F., Argnani, A., Trincardi, F., Bakran-Petricioli, T., Ceregato, A., Chimienti, G., Mačić, V., Poliseno, A. (2014). New Deep-water Cnidarian Sites in the Southern Adriatic Sea. *Mediterranean Marine Science*, **15**(2), doi: http://dx.doi.org/10.12681/mms.558.

[2] Armstrong, C. W., Grehan, A. J., Kahui, V., Mikkelsen, E., Reithe, S., Van den Hove, S. (2009). Bioeconomic Modeling and the Management of Cold-Water Coral Resources. *Oceanography*, **22**(1),86–91.

[3] Grehan, A.J., Unnithan, V., Olu-Le Roy, K., Opderbecke, J. (2005). Fishing impacts on deepwater coral reefs: making a case for coral conservation. In: Barnes, P.W., Thomas, J.P. (Eds.), *Benthic Habitats and the Effects of Fishing*. American Fisheries Society, Bethesda, MD, 819–832.

[4] Fahrmeir, L., Kneib, T., Lang, S. (2004). Bayesian Regularization in Structed Additive Regression: a Unifying Perspective on Shrinkage, Smoothin and Predictor Selection. *Statistics and Computing*, **20**(2), 203–219.

[5] Freiwald, A., Beuck, L., Rüggeberg, A., Taviani, M., Hebbeln, D. (2009). The White Coral Community in the Central Mediterranean Sea Revealed by ROV Surveys. *Oceanography*, **22**(1),36–52.

[6] Hastie, T.J., Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

[7] Ishwaran, H., Rao, J.S. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *Annals of Statistics*, 33(2), 730–773.

[8] Lin, X., Zhang, D. (1999). Inference in generalized additive mixed model using smoothing splines. *Journal of the Royal Statistical Society*, Series B, 61, 381–400.

[9] Scheipl, F. (2011). spikeSlabGAM: Bayesian Variable Selection, Model Choice and Regularization for Generalized Additive Mixed Models in R. *Journal of Statistical Software*, **43**(14), 1–24.

Statistical models for species richness in the Ross Sea

Cinzia Carota[1,*], Consuelo R. Nava[1], Irene Soldani[2], Claudio Ghiglione[3] and Stefano Schiaparelli[3]

[1] *Department of Economics and Statistics "Cognetti de Martiis", University of Torino; cinzia.carota@unito.it, consuelorubina.nava@unito.it*
[2] *aizoOn Technology Consulting; irene.soldani@aizoon.it*
[3] *DiSTAV, University of Genova and Italian National Antarctic Museum (section of Genova); claudio.ghiglione@riftia.eu, stefano.schiaparelli@unige.it*
[*] *Corresponding author*

Abstract. *In recent years, a large international effort has been placed in compiling a complete list of Antarctic mollusc distributional records based both on historical occurrences, dating back to 1899, and on newly collected data. Such dataset is highly asymmetrical in the quality of contained information, due to the variety of sampling gears used and the amount of information recorded at each sampling station (e.g. sampling gear used, sieve mesh size used, etc.). This dataset stimulates to deploy all statistical potential in terms of data representation, estimation, clusterization and prediction.*
In this paper we aim at selecting an appropriate statistical model for this dataset in order to explain species richness (i.e. the number of observed species) as a function of several covariates, such as gear used, latitude, etc.. Given the nature of data, we preliminary implement a Poisson regression model and we extend it with a Negative Binomial regression to manage over-dispersion. Generalized linear mixed models (GLMM) and generalized additive models (GAM) are also explored to capture a possible extra explicative power of the covariates. However, preliminary results under them suggest that more sophisticated models are needed. Therefore, we introduce a hierarchical Bayesian model, involving a nonparametric approach through the assumption of random effects with a Dirichlet Process prior.

Keywords. *Bayesian hierarchical model; Dirichlet Process; GAM; GLMM; Ross Sea.*

1 Introduction

Since many years, an international team of researchers has focused its attention on distributional data of Ross Sea (Antarctica) Mollusca, compiling a large dataset based on revised species identification and classification. The selection of this geographical position is crucial, especially in the light of the effects that climate changes might have on the biodiversity of the area. The dataset is the result of several scientific expeditions, performed with different goals, that span for a temporal timeframe of more than one century, specifically from 1899.

This dataset results to be highly asymmetrical in terms of available information. Expeditions in the last century essentially aimed at making a census of the Antarctic species while recent expeditions apply

balanced sampling designs that enable better statistical analyses and are focused on the study of species spatial and geographical distribution.

Hence, there have been some difficulties in the treatment and the adaption of the data collected before 2004, for instance due to the lack of information about species picked up dead or alive. Moreover, from 1899 to 2004, there is no record of "zero occurrences", i.e. stations that have been properly investigated but where no molluscs were found. This inevitably affects species richness, i.e. the number of different species observed in each sampling unit or station, which is the most used variable in biodiversity studies.

Despite these limitations, collected data remain a precious and unique source of information (see [8]) and several papers are going to be published based on these data, as [12]. Here we focus on species richness: our response variable Y. We also consider covariates such as the tools employed to collect sampling units. They can be grab, towed gears, Rauschert dredge (i.e. a towed dredge with a very fine mesh), or even "unknown" (i.e. where the gear was not recorded in the data log) or "multiple" (i.e. where more gears were deployed at the same station). In addition, geographical variables such as longitude, latitude, depth and distance from the nearest scientific station are taken into account. Successively, a factor geographical covariate, referred to as box[1], has been introduced.

2 Methods and results

We investigate the explanatory and predictive power of a large number of models and methods for count data. The simple Poisson regression model, inadequate because of over-dispersion, absence of zeros and excess of 1s in the data (see Figure 1 for a representation), has been variously enriched and made more flexible [3].

First, we introduce random effects of different nature alongside the effects of the covariates described in Section 1. In particular, we assume that $y_i \sim \text{Poisson}(\mu_i)$, for $i = 1,\dots,n$ and $log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \phi_i$, where $\mathbf{x}_i$ represents a $q \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed effects, and ϕ_i denotes a random effect accounting for observation specific deviations. In regarding the distribution of ϕ_i, denoted by G, two parametric assumptions are compared: $\phi_i \overset{iid}{\sim} N(0,\sigma^2)$ and $e^{\phi_i} \overset{iid}{\sim} \text{Gamma}(a,b)$ with $a = b$, so that $E(e^{\phi_i}) = 1$ and $Var(e^{\phi_i}) = 1/a$. The latter assumption introduces extra-variability on a different scale as ordinary predictors ([1], p.556) and leads to the Negative Binomial regression model [3, 9]:

$$y_i \sim \text{NB}\left(a, \frac{a}{a + e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right),\ i = 1,\dots,n. \tag{1}$$

where $E(Y_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}$ and $Var(Y_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}/a)$. As count data in ecology are often clumped (if the rate of capture of individuals varies randomly), producing an expected variance that is greater than the mean, in such literature [2] the parameter a is often referred to as the *clumping parameter* [2, 13].

We also explicitly consider special generalized linear mixed models, GLMM, where subsets of the n observations are given the same random effect, as for instance observations in the same box.

Given the absence of zeros, we explore truncated versions of the models just described and, for comparisons, we also apply linear mixed models to a log-transformation of the respose variable, a controversial practice very often recommended in the ecological literature.

Moreover, in stations where the gear is unknown, we try to impute its value in order to improve such a covariate.

Then, trying to increase the potential of all available covariates to explain the species richness, we explore more general parametric models such as generalized additive models, GAM [14].

[1] Boxes are defined with 1 degree of latitude and 1 of longitude. In the dataset 112 boxes are identified in which there have been at least the observation of one species.

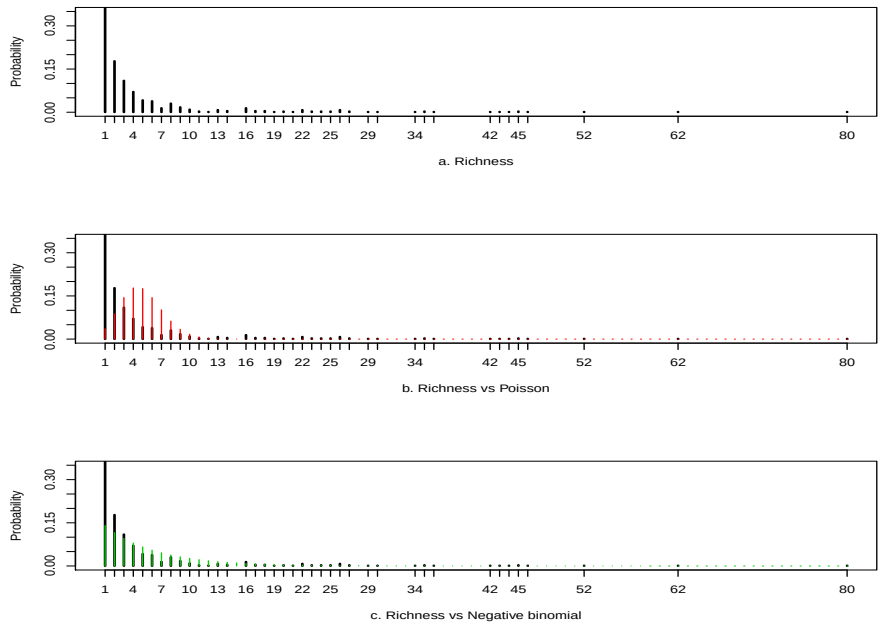


Figure 1: We compare the species richness, represented in plot **a**, respectively with $Y \sim \text{Poisson}(4.932)$ in graph **b** and with $Y \sim \text{NB}(0.977, 0.165)$ in plot **c**. The parameters of the Poisson and Negative Binomial distribution are estimated from the observed data.

Although, in terms of reduction of the residual deviance and AIC, all the strategies illustrated above provide appreciable contributions, their predictive power turns out to be further improvable, precisely because of the clumping of the data. In order to make the model able to capture the multimodal distribution of species richness (see Figure 1.a), we decided to re-interpret the described GLMMs as Bayesian hierarchical models and add the further level described below to the hierarchy.

We relax the assumption on the parametric form of the distribution function of random effects G and we model it by a Dirichlet Process prior $\mathcal{D}$ with base probability measure G_0 and total mass parameter m [7],

$$\phi_i | G \overset{iid}{\sim} G, \qquad G \sim \mathcal{D}(m, G_0),\ m > 0. \tag{2}$$

Considering that $E(G) = G_0$ and m controls the variance of the process, in practice G_0 specifies one's "best guess" about an underlying model of the variation in ϕ, and m identifies the extent to which G_0 holds ([6], p. 638). Within the class of models just defined, we consider specifications of G_0 that lead to direct generalizations of the GLMMs described above, namely $G_0 = N(\alpha, \sigma^2)$ and $G_0 = LG(a,b)$. LG denotes the distribution of ϕ_i, being $e^{\phi_i} \overset{iid}{\sim} \text{Gamma}(a,b)$ with $a = b$ as already discussed. Moreover, vague priors are assumed on β, a and m [4].

Under the previous assumptions, the likelihood function turns out to be a sum of terms where all possible partitions (clusterings) of the n observations into nonempty clusters are considered [10, 11]. This fact implies that:

i. to learn about a given observation/station, additional information to the one provided by covariates is borrowed from observations/stations belonging to the same subset, for each subset to which the observation can be assigned in the context of all possible partitions in nonempty subsets of the n

observations;

ii. the results under a hierarchical semi-parametric model with Dirichlet process random effects can be interpreted as averages over GLMMs, corresponding to all possible clusterizations of the $N(0,\sigma^2)$ or $LG(a,b)$ parametric random effects.

3 Conclusions

The natural implementation of discussed parametric statistical models – Poisson regression, Negative Binomial regression, GAM or GLMM – only partially explain our variable of interest. Multimodality and over-dispersion of species richness can be jointly modeled by adopting a more general non parametric hierarchical Bayesian approach as confirmed by the encouraging preliminary results we have obtained.

References

[1] Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

[2] Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** (3–4), 246–254.

[3] Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, vol. 53. Cambridge University press.

[4] Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015) Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Annals of Applied Statistics* **9**, 525-546.

[5] Clarke, A. (2008). Antarctic marine benthic diversity: patterns and processes. *Journal of Experimental Marine Biology and Ecology*, **366**(1),48–55.

[6] Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L. and Jordan, F. (2008). Modeling Unobserved Sources of Heterogeneity in Animal Abundance Using a Dirichlet Process Prior. *Biometrics* **64**, 635–644.

[7] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1**, 209–230.

[8] Griffiths, H.J., Danis, B. and Clarke, A. (2011). Quantifying Antarctic marine biodiversity: the SCAR-MarBIN data portal. *Deep-Sea Res II* **58**,18–29.

[9] Hilbe, J., (2007). *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK.

[10] Liu, J. S. (1996). Nonparametric Hierarchical Bayes via Sequential Imputations. *Annals of Statistics* **24**, 911–930.

[11] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Annals of Statistics* **12**, 351–357.

[12] Schiaparelli, S., Ghiglione, C., Alvaro, M. C., Griffiths, H. J., and Linse, K. (2014). Diversity, abundance and composition in macrofaunal molluscs from the Ross sea (Antarctica): results of fine-mesh sampling along a latitudinal gradient. *Polar biology* **37**(6), 859–877.

[13] Young, L.J and Jerry Youn, J. (1998). *Statistical Ecology*. Spinger.

[14] Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. and Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

Statistical analysis of zoo-agrarian crime

C. Cusatelli[1]*, M. Giacalone[2]

[1] *Ionian Department, University of Bari; carlo.cusatelli@uniba.it*
[2] *School of Economics, Management and Statistics, University of Bologna; massimilia.giacalone@unibo.it*
[*] *Corresponding author*

Abstract. *Environmental crime is a concept not easily defined as it necessarily encloses several, and dissimilar, type of offenses. Legislation too does not help as it lacks of a unique definition of environment. The current definition of "ecomafia", given by Legambiente a few years ago and now in the vocabulary of Italian language Zingarelli, includes a variety of criminal actions even in animals racket and in agriculture: the so-called "zoomafia" and "agromafia". The first flourishes on the control of illegal activities related to animals (illegal slaughter, cheating in horse shows, animal doping, theft thoroughbred, kennel business, fights between animals, illegal imports of puppies, poaching). The second affects, in Italy, a farmer out of three that are victim of threats, pressure and harassment, theft of equipment and agricultural vehicles or the commodities produced, theft of cattle for illegal slaughter and trade of meat, potentially dangerous to the health of consumers. Contrary to a mild and generalized decline in the number of offenses over the previous year, in 2013 both the agricultural sector, which has seen a surge of offenses (9,540: more than doubled), the waste cycle (5,025 crimes: +14.3%) and illegality committed against wildlife (8,504: +6.6%) recorded a growth.*

Keywords. *Ecomafia; Zoomafia; Agromafia.*

1 Ecomafia

Coined by Legambiente few years ago, the term Ecomafia indicates the activities of organized crime, involving environmental crimes. The data of 2013 show 29,274 offenses recorded, and the turnover of the environmental crime, always very high, despite the crisis, has reached almost 15 billion Euros within 321 surveyed clans. The slight decline in eco-criminal activity (in 2012 amounted to almost 16.7 billion) is due to the decline in investment at risk that also reduced earning opportunities for the gangs. Illegal trade in hazardous waste, amounting to 3.1 billion €, remains substantially unchanged, and sales of unauthorized construction is stable at 1.7 billion.

It is true that a cumbersome regulatory system, characterized by continuous emergencies and referrals, has in fact further expanded the range of all those who have experienced the possibility of easy enrichment at the expense of the environment and of the whole community.

2 Zoomafia

Born to control the illegal activities that have as their object the animals, the zoomafia is a crime association stretching from the north to the south of our country, involving the collaboration of the Italian organized crime with the foreign market. The turnover of the gangs specialized in this sector is estimated at 3 billion Euros: a round of huge money involving trade in dogs and cats with fake pedigree or exotic animals, poaching and smuggling of wildlife, illegal betting on street racing horses (one third of the total turnover), dog fights, fish racketeering, illegal slaughter, cattle rustling and adulteration.

In 2013 there were 8,504 offenses, increased by 6.6% compared to 2012. In particular, seizures increased from 418 (in 2012) to 2,620, and the arrests increased from 7 to 67: symptom that the repressive activity in the last year has been particularly effective. Sicily remains firmly in first place for number of offenses detected (with 1,344), followed by Campania (1,075), Puglia (953), Calabria (725) and Lazio (667). The top five provinces for the number of offenses against animals are Naples, Rome, Venice, Palermo and Trapani.

A growth in the number of crimes that is coupled with the cruelty and brutality with which it continues to do business on the skin of animals: clandestine races involving mostly drugged and abused horses, than slaughtered with contaminated meat put on the market; dogs used for fighting in improvised rings, including old vans, or used as carriers for drug shipments. It is estimated that in Italy there are about 100 thousand puppies illegally imported, usually coming from Hungary, Poland and Slovakia. There has been a parallel increase in seizures of illegal kennels, where many puppies are raised in cramped and unsanitary conditions.

The forms of business most profitable are connected to the world of horses and street racing: the Sos Impresa report argues that an illegal ride can produce up to 50 thousand Euros. This phenomenon is intertwined with other crimes such as the illegal slaughter: the world of horses is "polluted" by the interest of organized crime, and every year about 80 thousand horses arrive in Italy for slaughter, traveling in horrible conditions: salmonella, sunburn, stress syndrome are some effects of the long torture, with serious consequences for the quality of the meat and the health of unsuspecting consumers of horsemeat.

We must not forget the criminal behavior in relation to protected and endangered species: the trafficking of exotic animals, especially birds and their eggs, is becoming more and more prosperous. A growing traffic of wolves has been recorded in northern Europe, while Europol reports indicate that revenues of rhinoceros horns in Africa have become part of the budget of international terrorist organizations. Also the number of Italian customs controls showed an increase of illegal imports of species protected by CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora, also known as the Washington Convention): 5,485 animals seized in 2013 in Italy, where wildlife is threatened at various levels of risk, and monitoring is still limited: 20% of the national parks is subject to poaching, especially against the large mammals that inhabit them, such as deer, wild boar, chamois and roe deer.

In this context, that still lacks of a national legal framework to regulate the conservation and management of the entire fauna, the main threat to wild animals comes from a number of factors such as the apparent lack of political and social interest to monitoring wildlife and its conservation and management against the economic interest of unscrupulous people.

3 Agromafia

No sector of the economy and production seems to be overlooked by organized crime, and agriculture is one of the frontiers in the development of trafficking. The Agromafia is based on both investment

and laundering of money in crops, either by fraud to obtain public funds for the development of the agricultural sector: so impressive is the boom of the crimes in the food industry that from 4,173 crimes registered in 2012 rose to 9,540 with a doubling of complaints and 57 people arrested. Mafia is directly involved in the entire chain: from the field, through transporting, to the markets, and often this continues laundering money through investments in hotels, restaurants and pizzerias. Especially in the regions of southern Italy, there are thousands of producers who are subject to threats, harassment and extortion. The countryside, then, is a world in which, compared to urban areas, it still retains very strong code of silence with respect to this type of domain.

Also in this area it is needed to improve the efforts of the police, strengthening of investigations, on the one hand, favoring the short chain and quality agriculture on the other. There is talk of cattle rustling that today feeds a chain of illegal animals without health checks, trade and use of veterinary medical substances not permitted, often made without compliance with the minimum hygiene rules, trade of meat products of poor quality, potentially dangerous to the health of consumers.

The fourth report on crime in agriculture, prepared by the Italian National Confederation of Farmers (CIA) in collaboration with the Foundation Humus, says that agriculture generates income for the "company Mafias Corporation" for more than 50 billion Euros a year, equivalent to just under a third of the illegal economy in our country (169.4 billion Euros). A bargain widespread throughout the national territory that is meanly speculating on the difficulties caused by the economic crisis: 25 thousand companies forced to close because of organized crime, 150 thousand head of cattle that disappear, 350 thousand farmers victims of racket, lace, extortion and aggression.

Monitoring agricultural lands means managing some prominent productions of our agri-food sector, so even aspire to be the beneficiaries of public funding to support the economy of the southern regions classified Convergence Objective (European Structural Funds). And if some gangs are characterized by historic control of the fruit and vegetable markets, transport and distribution of products, other criminal holdings have specialized in adulteration and in counterfeiting of trademarks and the so-called Italian sounding. According to the tenth dossier "Italy at the table in 2013" by Legambiente and the Movement of the Citizen Defense, in that year 500,000 inspections were carried out and 28,000 tons of products were seized, for an economic value of over half a billion Euros. Nothing unusual for an industry that handles annually about 245 billion Euros between consumption, export, distribution and induced: about 15% of the national GDP. According to estimates by the Ministry of Agriculture and Forestry, counterfeiting produces in Italy more than 4 billion Euros, while in the rest of the world the false "Made in Italy" has a value of about 50 billion Euros. Money accumulated plundering the wealth of our country is based on cheating especially healthy companies, i.e. those who respect the law trying with great difficulty to do their part.

4 Final remarks

All listed problems, which increase over the years, have not only a political or economic issue, in Italy. Our drama is a baffling and general degradation of ethics that is poisoning our Earth. It should therefore be defined a real national plan to combat Ecomafia, in consultation with the regions and local authorities, to prepare all the tools to control the territory. Unfortunately the penalties for environmental crimes continue to be almost exclusively fines, and jail time for the perpetrators of these crimes remains sometimes a pure utopia. In recent years there has been a further twist: our Parliament, in the name of simplification, decriminalized offenses against animals for so-called tenuous nature of the fact.

In conclusion, we hope that, in compliance with the requirement of legality and safety widespread in all strata of society, some laws waited for years for the protection of animals are implemented. We refer also to the amendment of legislation on companion animals and the rules on the protection from attacks by dogs, which require restoration in a single organic and renewed text.

References

[1] Angelin, A. (2004). *La società dell'ambiente*. Armando Editore. Roma.

[2] Bottino, G. (2008). *Codice dell'ambiente, Edizione 101*. Giuffrè Editore. Milano.

[3] Legambiente, (2014). *Osservatorio Ambiente e Legalità, Ecomafia 2014, Le storie e i numeri della criminalità ambientale*. Edizioni Ambiente. Milano.

[4] Pierobon, A. (2012). *Nuovo manuale di diritto e gestione dell'ambiente. Analisi giuridica economica, tecnica, organizzativa*. Maggioli Editore. Santarcangelo di Romagna.

[5] Pollifroni, M. (2010). *Green public accounting. Profili di rendicontazione ambientale per un'azienda pubblica responsabile e sostenibile*. G Giappichelli Editore. Torino.

[6] Vagliasindi, G.M. (2012). *Attività d'impresa e criminalità ambientale. La responsabilità degli enti collettivi*. Libreria Editrice Torre. Catania.

Avoiding the global change in climate

V. Demchuk[1,*] and M. Demchuk[2]

[1] *The Department of Physics and Technology, Rivne State Humanitarian University, Rivne, 31 Ostafova street, 33028, Ukraine; demchukvb@ukr.net*
[2] *ACADEMIA-RESEARCH.COM, Ukraine; nbdemch@gmail.com*
[*] *Corresponding author*

Abstract. *Nowadays, humankind extracts most of the energy it consumes from fossil fuels. Unfortunately, it entails building up the carbon dioxide in the atmosphere. People have to cut the emissions of this gas drastically to avoid the global change in the climate in few decades. One of the options they have consists in using solar energy as a primary energy source. In this case, the main difficulty consists in the large scale energy storage needed for satisfying needs in energy at night. In this work, we give an example of energy storage in polymer surface layers of a polymer matrix composite filled with magnetic microparticles. Specifically, in the recent research V. B. Demchuk argued that improvement of mechanical properties of Polyvinyl chloride (PVC) composite materials filled with ferrite microparticles reached at sufficiently high filler concentrations through influencing the formation of these microcomposites by constant magnetic fields (CMFs) occurs thanks to the fact that PVC macromolecules situated in the polymer surface layers of these microcomposites are pushed out of regions of high magnetic field intensity. His reasoning is based on integrated analysis of laboratory measurements and numerical calculations. In this paper, we present computational abstractions that allow handling the chaotic disposition of the filler particles during calculation of magnetic fields in polymer surface layers of microcomposites near phase transition points under the constraint due to limited performance of modern computers.*

Keywords. *Global change in climate; Polymer surface layer; Phase transition; Integrated analysis, Chaotic disposition.*

1 Introduction

Cutting emissions of the carbon dioxide drastically is a top priority for humankind because building up this gas in the atmosphere causes global changes in the climate. They can become irreversible if nothing will be done regarding this matter during next few decades. This problem is challenging since people extract energy mainly from fossil fuels. Unfortunately, scientists have not come up with the technology of large scale carbon capturing yet. Moreover, such the technology entails difficulties with the storage of the extracted gas. Specifically, large quantities of the carbon dioxide cannot be stored in oceans due to ecological considerations. Besides, they cannot be stored underground since the leakage of this gas puts in danger human lives. Therefore, significant effort should not be put in this direction. Solar energy is the source that is rich enough to become a new primary energy source for the humankind. Although this source is not available round the clock, potentially it can satisfy all human needs in electricity many

times over. Therefore, to switch to extracting most of the needed energy from this source, people have to come up with the technology of the large scale energy storage [1]. Each year the amount of energy that is stored in the result of photosynthesis is enough to satisfy all human needs in energy several times [2]. Hence, chemical bonds are a proper place to store energy. To avoid additional expenditures, it is important to discover examples of the energy storage in chemical bonds of the industrially mastered materials. In this work, we provide an example of such the storage in polymer surface layers of polymer matrix composites (PMCs) containing only those polymers that are mass-produced. In Section 2, we present this example and describe the research methodology that led to its discovery in the reference [3]. In Section 3, we illustrate one of the elements of this methodology. Section 4 contains a brief summary.

2 Storing energy in polymer surface layers of PMC

Mechanical properties of an industrially mastered polymer filled with magnetic microparticles can be improved through influencing the formation of the composite with a constant magnetic field if the filler concentration is higher than the one at which the phase transition occurs [4]. There are two interface phenomena that can be responsible for this improvement. Specifically, a constant magnetic field can orient macromolecules of the polymer surface layer and these molecules can be pushed out of the regions of the high magnetic field intensity [3]. In the reference [3], V. B. Demchuk argues that it is the second phenomenon that is responsible for the above mentioned effect. Although he uses classical considerations, they are relevant since the classical physics is the limiting case of the quantum one. Both these phenomena assume energy storage in chemical bonds. Traditional numerical analysis of real life problems assumes parallel computing and is hindered by various scale effects induced by a choice of boundary conditions. Unfortunately, these effects become extremely important near phase transition points [5]. Therefore, in the paper [3], to figure out which interface phenomenon is responsible for the improvement of the mechanical properties of the microcomposite, the Cyber-Physical System (CPS) methodology is used. In a CPS, computing experiences the lack of resources and is integrated with physical processes. Finding new computational abstractions is important for development of this emerging field [6]. Such the generalization of notions is essential of the framework of computational thinking. For example, the abstraction of an algorithm is not supposed to produce the desired output within a finite modern processor time frame [7]. The following section presents computational abstractions that allow calculating magnetic fields in polymer surface layers of microcomposites near phase transition points.

3 Handling chaotic dispositions of magnetic particles during calculations of the magnetic fields

In the recent research [4], to prepare PMCs, PVC was mixed with Fe_3O_4 fine powder. The mixture was exposed to pressure and temperature as high as respectively 10 MPa and 420 K and to the external magnetic field so strong that all powder particles were in the state of the magnetic saturation. Taylor's theorem allows us to suppose that calculating the magnetic field in the vicinity of the filler particle situated in the PMC we can neglect the influence of the other Fe_3O_4 particles. Therefore, in the model # 1 the calculation is performed for a single spherical Fe_3O_4 particle with the uniform magnetization density $\vec{M}$ placed in the external CMF $\vec{H}_0$ (see Fig. 1). However, the size of the surface layer can turn

out to be large enough for the other filler particles to influence the magnetic field inside of it significantly. In this case, this problem cannot be solved precisely due to randomness of filler particle disposition in a PMC. Therefore, it can be referred to the class of artificial intelligence problems and should be solved using simplifying assumptions. In the recent research [4], PMC samples were prepared

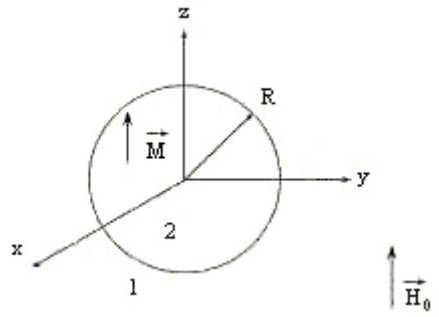


Figure 1: Set up of the model # 1.

in the form of cylinders with their diameters equal to $2.5 \cdot 10^{-2}$ m and their heights equal to $5 \cdot 10^{-3}$ m. Since the sample height is 5 times smaller than its diameter, in the model # 2 the boundary effects are neglected and it is assumed that the Fe_3O_4 particles create the quasi lattice in the 3-dimensional space as shown on Fig. 2. Calculations according to the model # 3 are performed to estimate the error of

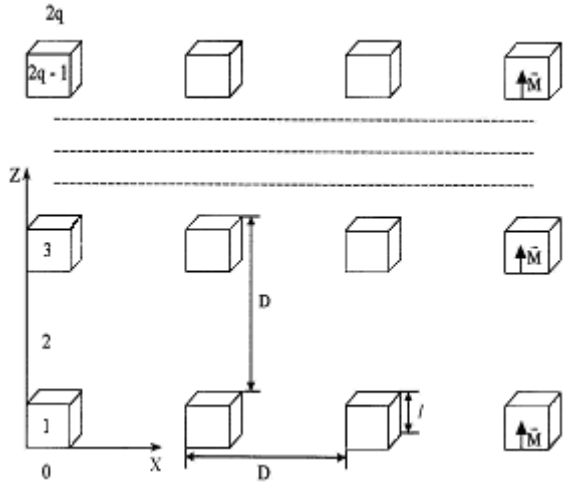


Figure 2: Filler particle disposition assumed in the model # 2.

magnetic field calculation according to the model # 2 that arises due to the substitution of the disposition of filler particles shown on Fig. 2 for the chaotic one in the PMC in the framework of the model # 2. Therefore, the filler particle disposition in the model # 3 is the quasi lattice that can be obtained from the quasi lattice of the model # 2 through shifting filler particle layers denoted on Fig. 2 by numbers $4 \cdot p - 1$ where $p = \overline{1, \widetilde{p}}$, $\widetilde{p} = q/2$ if q is even and $\widetilde{p} = (q-1)/2$ if q is odd (here and below q is the number of horizontal filler particle layers in the quasi lattice of the model # 2) at the distance equal to the half of the lattice period in the direction of the x-axis with subsequent shifting them at the same distance in the direction of the y-axis. Such the assumptions about dispositions of filler particles allow applying Fourier's method to the calculation of the magnetic field in the frameworks of the models # 2 and # 3.

4 Conclusions

During next few decades people have to make radical changes in the way they extract energy. They cannot keep extract energy mainly from fossil fuels even if they come up with the technology of the large scale carbon capturing since the extracted gas cannot be stored in oceans and underground due to respectively ecological and safety considerations. Humankind should turn its attention to solar energy. This source is so rich that all human needs in energy are only its tiny portion. However, people cannot take advantage of this source at night. Therefore, to start using solar energy heavily people have to come up with the technology of large scale energy storage. In this work, it is pointed out that such the storage can be realized in chemical bonds of mass-produced materials. Specifically, energy can be stored in polymer surface layers of industrially mastered polymers filled with magnetic microparticles [3]. This effect was discovered thanks to integrated analysis of the results of numerical modeling and the ones of experimental measurements. In this work, we present the computational abstractions that allow handling random disposition of the magnetic microparticles inside a polymer composite filled with magnetic microparticles near a phase transition point during calculation of a CMF in a polymer surface layer of the composite.

References

[1] Lewis, N. S. (2009). "The Roger Revelle centennial symposium series: Powering the planet." *YouTube*. <http://www.youtube.com/watch?v=f1sYmBX7rNA >.

[2] Uner, D. (n. d.). Storage of chemical energy and nuclear materials, *Energy Storage Systems* **2**, *in Encyclopedia of Life Support Systems (EOLSS)*. Eolss Publisher. Paris. < http://www.eolss.net/sample-chapters/c08/e3-14-05.pdf >.

[3] Demchuk, V. B. (2014). Influence of constant magnetic fields on formation of polymer surface layers in polymer matrix microcomposites. *Chem. Rapid Commun*. **2 No 3**, 48-54.

[4] Demchuk, V. B., Kolupaev, B. B., Klepko, V. V., and Lebedev, E. V. (2012). Influence of external magnetic field on intrinsic pressure of the system PVC-magnetite. *Physics and Technics of High Pressures* **22(2)**, 95–109. <http://www.fti.dn.ua/site/ftvd-journal/journal-content/ftvd-v22-2/>.

[5] Allen, M. P. and Tildesley, D. J. (1991). *Computer Simulation of Liquids*. Oxford University Press. New York.

[6] Shi, J., Wan, J., Yan, H., and Suo, H. (2011). A survey of Cyber-Physical Systems. *In Wireless Communications and Signal Processing (WCSP), 2011 International conference on*. <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6096958>.

[7] Wing, J. M. (2008). Computational thinking and thinking about computing. *Phil. Trans. R. Soc. A.* **366 (1881)**, 3717–3725. <http://rsta.royalsocietypublishing.org/content/366/1881/3717.full>.

Modeling cement distribution evolution during permeation grouting

M. Demchuk[1,*] and N. Saiyouri[2]

[1] *ACADEMIA-RESEARCH.COM, Ukraine; nbdemch@gmail.com*
[2] *Institute of Mechanics and Engineering (I2M), Bordeaux, France; nadia.saiyouri@u-bordeaux1.fr*
[*] *Corresponding author*

Abstract. *Permeation grouting is a proper technique for strengthening dry sand before a tunnel construction in it. In the fifties of the past century, N. N. Verygin modeled this technique with 1-dimensional problems formulated in domains with a free moving boundary and came up with their analytical solutions. As for the 2-dimensional set ups, respective problems are formulated in domains that have complicated shapes and contain free moving boundaries. Until recently these difficulties seemed to be insuperable and all 2-dimensional grouting models in the framework of the continuum approach were based on the convective dispersion equation. They can be classified as the ones that describe pollution propagation. Nevertheless, M. B. Demchuk has lately come up with numerical solutions of 2-dimensional problems with free moving boundaries which set ups correspond to in situ grouting. Moreover, he has shown the following: adoption of the continuum approach is relevant for the set of input parameters used, among the curvilinear grids the calculations are performed on there are the ones that have chaotic dispositions of their nodes in space on some time layers. In this work, rough estimates are performed that indicate that the use of problems with free moving boundaries in the numerical modeling in hand is relevant.*

Keywords. *Permeation grouting; Free moving boundary; 2-dimensional models; Chaotic disposition of nodes; Pollution propagation.*

1 Introduction

Strengthening dry sand must precede the tunnel construction in it to have enough time for installing temporary support. In this case, to reduce the total cost of excavation it is desirable to preserve the structure of the treated soil. Therefore, permeation grouting is a proper technique for such the soil reinforcement. It assumes injecting cement grout in a treated soil through an injector. The quality of the soil reinforcement significantly influences the overall efficiency of the construction. This technique is rather costly and time consuming. The cement concentration distribution evolution determines the regime of the grouting [1]. Therefore, its mathematical modeling is important. There are a lot of papers devoted to modeling permeation grouting. The models [1]–[3] are based on the convective dispersion equation and can be classified as the ones that describe pollution propagation [4]. In the paper [5], this technique is modeled with a 1-dimensional problem containing a free moving boundary. In the recent paper [6], the last approach is generalized for the cases of 2-dimensional set ups that correspond to *in situ* permeation grouting. Specifically, in the research [6], M. B. Demchuk performed calculations on sparse curvilinear spatial grids and estimated the truncation errors of the final injection front position calculations neglecting the uncertainty in the final injection front position due to uncertainty in the choice

of the method of the injection front interpolation on every time layer. In the paper [7], M. B. Demchuk and O. G. Nakonechnyi conducted the numerical experiment that verified the validity of this assumption. Moreover, its results indicate that the curvilinear grids the calculations [7] are performed on have chaotic dispositions of their nodes in space on some time layers. The aim of this work is to show that the use of problems with free moving boundaries in the numerical modeling [6] is relevant.

2 Modeling permeation grouting with problems containing a free moving boundary

The filtration of a cement grout in a soil is an example of a creeping flow. Describing such the motion, one can neglect the inertia forces in comparison with the frictional ones. The cement grout solution becomes more homogenous when the grain concentration increases [1]. Therefore, filling partially pores with it during its injection in a dry soil can be modeled as its dissolution in a fictitious weightless fluid with zero viscosity that saturated the soil before the injection. In the paper [4], M. B. Demchuk analyzed the analytical solution of the following partial differential equation:

$$D\partial^2 c/\partial x^2 - V\,\partial c/\partial x = \partial c/\partial t \qquad (1)$$

with such the initial and boundary conditions:

$$c(x,0)=0,\ c(\infty,t)=0,\ c(0,t)=\hat{c} \qquad (2)$$

where D, V, and $\hat{c}$ are positive constants. He showed that the injection front is sharp only if

$$V\cdot t >> \sqrt{8\cdot D\cdot t}\ . \qquad (3)$$

If V is the velocity of particles of the fluid phase, then injection implies

$$D \approx a_L \cdot V \qquad (4)$$

where a_L is the coefficient of the longitudinal dispersion. From equations (1)–(4), it follows that describing injection of a cement solution in a fluid saturated soil one can neglect the peculiarities of a solute propagation in a porous medium and model the cement concentration distribution evolution with a problem containing a free moving boundary if the following holds

$$\sqrt{V\cdot t} >> \sqrt{8\cdot a_L}\ . \qquad (5)$$

The phenomenon of the mechanical dispersion taking place during cement grout solution injection in a fluid saturated soil ($a_L V >> D^*$ where D^* is the diffusion coefficient) occurs due to the gradual change of the velocity of the fluid particles from the zero value on the surface of the pore to the maximal one in some internal point of the pore and due to the fact the pores in the ground can be viewed as a chaotic system of interconnected tunnels [8]. Comparison of injection of cement grout solution in water saturated sand with the one in the dry sand when other conditions are the same suggests that in the last case the ratio of the number of grains that are situated in a zone of the transition from the soil with the maximal cement concentration to the one with the zero concentration and simultaneously are in the contact with the pore surface to the total number of grains situated in the zone is greater than such the ratio in the first case. Hence, from dimensional considerations it follows that a_L for the fictitious fluid saturated sand is smaller than a_L for the respective water saturated one. Therefore, checking the validity of the condition (5) for the dry sand, one can substitute in its right hand side a_L for the respective water saturated sand.

3 Numerical estimations

In the paper [6], four problem set ups are considered. In the cases of the set ups # 1 and # 3, it is assumed that a long trench is made under an injector foundation. Its width is $2 \cdot r_0$ and its depth is h_0. The astringent infiltrate is injected in this trench at the constant pressure p_0 (see Fig. 1). In the set ups #

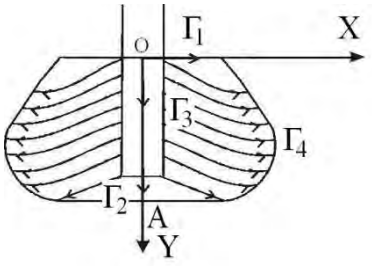


Figure 1: Problem set ups.

2 and # 4 we have the round bore-hole instead of the trench and assume that other conditions are the same. Its radius is r_0 and its depth is h_0. In the set ups # 1 and # 2, the ground skeleton is regarded as absolutely rigid whereas in the other ones it is assumed to be deformable. In each case $h_0 >> r_0$, the injection front (the curve Γ_4 on Figure 1) is a free surface, and its evolution in time and space needs to be found. In the recent research [6], the truncation error of the final injection front position calculation is so large that final free surfaces obtained for the absolute rigid ground skeleton can be compared with the respective ones obtained for the deformable ground. On Figure 2, the evolution of the injection front during permeation grouting corresponding to the case of the set up # 4 and the softest ground obtained in the numerical modeling [6] is presented. From this figure, it follows that the point of the intersection of the injection front with the vertical axis (the point A on Figure 1) moves slower than other points of the free surface. It covers the distance equal to 1.193 m. In the modeling [6], the treated soil is sand. Since for the water saturated sand $a_L = 1 \cdot 10^{-2}$ m [3], the condition (5) is fulfilled for all evolutions obtained in [6].

4 Conclusions

In this work, rough estimates are performed that indicate that problems with free moving boundaries are properly used to model the cement concentration distribution evolution during permeation grouting in the recent research [6].

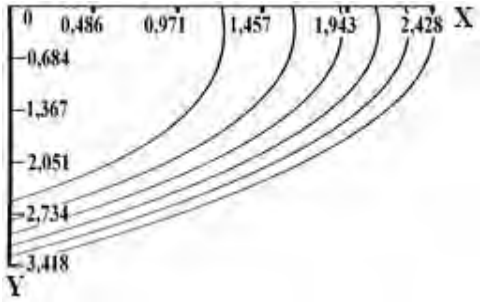


Figure 2: Injection front evolution [4].

References

[1] Bouchelaghem, F. and Vulliet, L. (2001). Mathematical and numerical filtration-advection-dispersion model of miscible grout propagation in saturated porous media. *International Journal for Numerical and Analytical Methods in Geomechanics* **25**, 1195–1227.

[2] Chupin, O., Saiyouri, N., and Hicher, P.-Y. (2008). The effects of filtration on the injection of cement-based grouts in sand columns. *Trans. Porous Med*. **72(2)**. 227-240.

[3] Chupin, O., Saiyouri, N., and Hicher, P.-Y. (2009). Modeling of a semi-real injection test in sand. *Computers and Geotechnics* **36(6)**, 1039-1048.

[4] Demchuk, M. B. (2010). Mathematical modelling of a process of injection of an astringent grout in a porous medium. *Mathematical and Computer Modelling. Series of Physical and Mathematical Sciences* **4**, 61–75.

[5] Verygin, N. N. (1952). Injection of astringent solutions in rocks with the aims of improving strength characteristics and watertightness of the bases of waterside structures. *The News of the Academy of Sciences of the USSR the Department of Technical Sciences* **5**, 674-687.

[6] Demchuk, M. B. (2013). Adoption of the continuum approach in real scale grouting models. *Mathematical Machines and Systems* **3**, 170–177.

[7] Demchuk, M. B. and Nakonechnyi, O. G. (2013). Injection front interpolations in real scale grouting models. *Transactions on Computer Systems and Networks (Lviv Polytechnic National Press)* **773**, 165—178.

[8] Bear, J. and Bachmat, Y. (1990). *Introduction to Modeling of Transport Phenomena in Porous Media.* Kluwer Academic Publishers. Dordrecht.

Patterns and Processes Revealed in High-Frequency Environmental Data

A. Elayouty[1,*], M. Scott[1], C. Miller[1] and S. Waldron[2]

[1] *School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; a.el-ayouti.1@research.gla.ac.uk, Marian.Scott@glasgow.ac.uk, Claire.Miller@glasgow.ac.uk*
[2] *School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK; Susan.Waldron@glasgow.ac.uk*
[*] *Corresponding author*

Abstract. *High-frequency data are informative but also very challenging to analyze. Appropriate statistical tools are required to extract useful information from such data. A 15-minute resolution sensor-generated time series of the $EpCO_2$ from October 2003 to August 2007 in a small order river system in Scotland is used as an illustrative dataset. The aim of this paper is to study the daily patterns and dynamics of $EpCO_2$ using a Functional Data Analysis (FDA) approach. Using FDA, the discrete data within each day have been transformed to a smooth curve; then, a K-means clustering procedure has been applied to the spline coefficients defining the daily curves to identify the common daily patterns which can then be linked to underlying climatological and hydrological conditions.*

Keywords. *High-Frequency Data; Partial Pressure of Carbon Dioxide, Functional Data Analysis.*

1 Introduction

Advances in sensor technologies enable environmental monitoring programmes to record and store data at high temporal frequencies. These technical improvements in data acquisition present an opportunity to improve our understanding of environmental systems. However, to benefit from this wealth of data, appropriate statistical tools are required to manipulate and analyze large volumes of serially correlated data. In this paper, we consider a 15-minute resolution sensor-generated time series of the over-saturation of CO_2, $EpCO_2$, from October 2003 to August 2007 in a small order river system of the River Dee, Scotland. Surface waters are considered as key sources of atmospheric CO_2, therefore comprehensive understanding of the CO_2 dynamics in surface waters, quantified by the $EpCO_2$, is important. Due to the high-frequency nature of the data and the complex dynamics of $EpCO_2$ in relation to hydrodynamics, sophisticated exploratory tools and statistical models are needed to extract the main characteristics of the $EpCO_2$ series. One approach to analyze the high-resolution $EpCO_2$ time series is to investigate and model its variations and relationship with hydrology over time using wavelets and additive models (see [2] for details). Another strategy is to consider a functional data analysis approach, which is the main focus of this paper.

In Functional Data Analysis (FDA) [4], a time series can be treated as observations of a continuous function collected at finite series of time points, the observations of interest for data analysis are then curves over time. The paper describes the analysis of the daily dynamics of $EpCO_2$ using an FDA approach. In particular, the ultimate goal of the paper is to investigate the common daily patterns of $EpCO_2$ based on both mean level and shape, using functional clustering techniques. This, in turn, will help in determining the underlying climatological and hydrological conditions responsible for the different $EpCO_2$ daily patterns.

2 Methodology

In the context of FDA, the 96 (15-minute) observations within each day are considered as the discrete observations of a continuous smooth function. This view of the data allows the daily $EpCO_2$ patterns to be estimated using smooth curves removing issues of high correlations and variability between 15-minute observations. In this setting, the observations of interest are daily curves or functions, which are considered as realizations of a functional stochastic process $(X_i(t) : i \in \mathbb{Z})$, such that the time parameter i is discrete representing day of the year and the time parameter t is continuous representing time of the day. That is, $x_i(t)$ is regarded as the observation on day i, with intraday time parameter t.

Using the `fda` package in `R`, a smooth curve $x_i(t)$ is fitted for the observations within each day $(x_{i1},\ldots,x_{i96})$, $i = 1,\ldots,1095$ using cubic B-splines combined with a second-order roughness penalty, such that $x_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t) = \mathbf{a}_i^T\Phi(t)$, where $\mathbf{a}_i$ is the vector of basis coefficients $(a_{i1},\ldots,a_{iK})^T$ to be estimated for the i^{th} sample path using penalized regression splines and $\Phi(t)$ is the vector of the basis functions $(\phi_1(t),\ldots,\phi_K(t))^T$.

With analogy to any classical statistical analysis, detecting outliers is crucial. Functional outliers can be identified using functional boxplots [5], developed based on the "band depth" measure which determines how deep or central a curve is. As with classical boxplots, functional boxplots are then constructed and functions are flagged as outliers if they fall outside the boxplot fences obtained by inflating the interquartile range (IQR) by $1.5 \times$ IQR. According to [5], functional boxplots are able to detect both shape and magnitude outliers (see [5] for more details).

After removing the detected outliers, a functional clustering procedure is performed to visualize the similarities and differences between the daily $EpCO_2$ curves and highlight the underlying climatological and hydrological conditions. One approach is to cluster the daily curves based on their spline coefficients using classical clustering techniques such as K-means [1]. The K-means procedure is iterative, in which the number of clusters is first specified, then each object is assigned to the cluster with the nearest mean such that the within-cluster sum of squares is minimized. The optimal number of clusters is initially selected using the gap statistic [6], which compares the change in the observed within-cluster dispersion with that expected under a null reference distribution of no clustering.

3 Results

Initially, a smooth curve is fitted for the observations within each day using saturated cubic B-splines combined with a roughness penalty. The smoothness of the curves is controlled by the smoothing parameter selected based on a sensitivity analysis. Next, functional boxplots were used to detect the out-

lying daily curves. The $EpCO_2$ curves of 29/8/2005, 23/9/2005 and 27/9/2005 are marked as shape and magnitude functional outliers and 26/9/2005 is flagged as a magnitude outlier. It is unsurprising to find outliers in adjacent days since the daily curves are time dependent. This number of potential outliers might decrease if the correlation between the curves is taken into account. However, 4 functional outliers represent only 0.4% of the total number of curves, and hence it was decided to delete these 4 curves before proceeding with any further analysis.

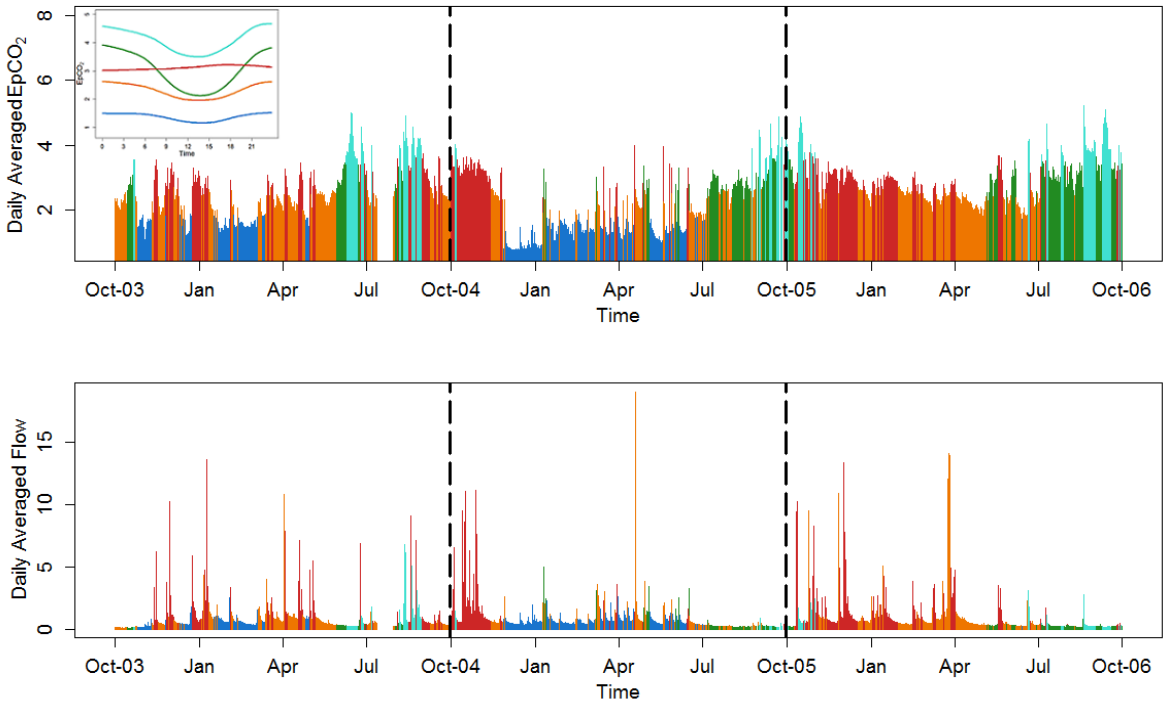


Figure 1: The average daily $EpCO_2$ (top) and flow (bottom) colored by their class membership obtained using K-means clustering of the $EpCO_2$ daily curves' splines coefficients. The top left legend shows the mean curve of each cluster.

After summarizing each daily curve with its estimated spline coefficients, K-means clustering was applied for a range of different number of clusters and the gap statistic was calculated to select the optimal number of clusters. The gap statistic identified consistently 5 optimal groups to represent the daily patterns of $EpCO_2$ for curves fitted using different smoothing parameters. This indicates that the optimal number of clusters is not sensitive to the chosen smoothing parameter. Next, a K-means procedure with 5 centers is applied to the spline coefficients. The top left legend in Figure 1 shows the grouping structure of the $EpCO_2$ daily curves which clearly depends on the estimated $EpCO_2$ mean levels. This is because the clusters are formed using the K-means procedure in which the classification is primarily based on the mean level. Another element of distinction between the 5 groups is the shape of the daily pattern of $EpCO_2$. Some of the groups have a clear daily cycle with a drop in the $EpCO_2$ during day time while others have a fairly constant $EpCO_2$ level over the day. The top and bottom panels of Figure 1 display the daily average $EpCO_2$ and flow (indicative for wet/dry days) respectively, and the class membership of each day according to the results of K-means clustering of the $EpCO_2$ daily curves. The figure shows that the turquoise and green curves representing a generally high $EpCO_2$ average with medium to severe drops in the $EpCO_2$ average level during the day light hours are more prominent in dry summer days, whilst the orange curves characterized by a lower $EpCO_2$ average and a medium trough during daytime underly the relatively wet spring and summer days. The blue curves with fairly stable and relatively

low levels of average $EpCO_2$ characterize the dry periods of winters and springs and the red group of curves consisting of a variable set of daily patterns often corresponds to the high flow events occurring in autumn and winter.

4 Discussion

In conclusion, FDA is shown to be a key tool in analyzing high-frequency environmental data. FDA has allowed the data observed every 15 minutes within each day to be expressed as continuous smooth functions without being concerned about the high-correlations between the 15-minute observations within the same day. After detecting the functional outliers, the primary results of functional clustering analysis indicated that the mean $EpCO_2$ level underlies the grouping structure. It is also evident that the $EpCO_2$ daily pattern is determined partly by the underlying hydrological (flow) and climatological (season) conditions.

Further work will investigate classifying the daily curves based on their Functional Principal Components scores (FPCs) using the classical clustering algorithms. Two key advantages for the use of FPCs are (i) identifying the primary sources of variations in the daily patterns of $EpCO_2$ and; (ii) the orthogonality and hence the independence between the FPCs of the same smooth curve. The shortcoming of either functional clustering approach described here is that the serial dependence between the daily curves has not been taken into account. Therefore, current work involves the extension of dynamic FPCs [3] which take advantage of the serial dependence between curves.

Acknowledgments. AE is grateful to the Glasgow University sensor studentship for funding.

References

[1] Abraham, C., Cornillon, P., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics* **50**, 581–595.

[2] Elayouty, A., Scott, M., Miller, C., Waldron, S. and Franco-villoria, M. (2015). Challenges in Modeling Detailed and Complex Environmental Data Sets: A Case Study Modeling the Excess Partial Pressure of Fluvial CO_2. *Journal of Environmental and Ecological Statistics* **Manuscript under revision**.

[3] Hormann, S. and Kidzinski, L. (2014) Dynamic Functional Principal Components. *Journal of Royal Statistical Society* **77**, 319–348.

[4] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*. Springer.

[5] Sun, Y. and Genton, M. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics* **20(2)**, 316–334.

[6] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society* **63(2)**, 411–423.

Multivariate spatio temporal models for large datasets and joint exposure to airborne multipollutants in Europe.

A. Fassó[1,*], F. Finazzi[1] and F. Ndongo[1]

[1] *University of Bergamo, Via Marconi 5, 24044 Dalmine BG, Italy; alessandro.fasso@unibg.it*
[*] *Corresponding author*

Abstract. *We consider the distribution of population by exposure to multiple airborne pollutuants at various spatial and temporal resolutions over Europe. The estimation of this distribution and its uncertainty are obtained via model based high resolution semiparametric estimates of daily average concentrations for seven pollutants in years 2009-2011. In order to exploit the spatial information content and allow the computation of daily multipollutant exposure distribution, uncertainty included, we use a multivariate spatio-temporal model capable to handle non Gaussian large datasets such as multivariate and multiyear daily air quality, land use and meteorological data over Europe.*

Keywords. *EM algorithm; D-STEM software; Air quality.*

1 Introduction

Following the development of modern fixed monitoring networks, human exposure to airborne pollution is often related to the so called "ambient exposure", namely the pollutant concentration at which people are exposed when outdoor, which is assessed by means of pollutant concentration data coming from the mentioned monitoring networks. Since using the ambient exposure as personal exposure may lead to an ecological fallacy, individual exposure is deserving an increasing attention in recent years. The exposure of an individual to airborne pollutants can be defined as the total or average amount of pollutants the person is exposed to (and can breathe) during a given period of time of his/her life. Clearly, this amount depends on a large number of factors, the main factors being the spatial location of the person and, conditionally on this, his/her life style and activity level. This is often called "personal exposure" as, in principle, it can be related to each specific person.

In the recent statistical literature, two main approaches to model personal exposure emerged. On the one side the probabilistic approach of Zidek et al. (2005) allows to simulate individual exposure temporal profiles for a finite set of individuals and, based on this, it allows to estimate a simulated population distribution. On the other side, an empirical approach based on personal monitors has been considered be few authors, see e.g. McBdride et al (2007) or Jahan et al (2013).

In principle, aggregating personal exposures bring us to population exposure of an entire city or country or continent. Since this *total* personal exposure is difficult or impossible to compute in practice,

Pollutant	C_6H_6	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
Stations	344	591	1978	1800	1837	748	1340
Missing	47%	30%	22%	16%	23%	32%	26%

Table 1: European monitoring network details.

Finazziet al. (2011) and Shaddick et al (2013) consider a simplified concept of "population exposure", or exposure burden, which is intermediate between ambient and personal exposure and assumes that, for a large fraction of the population, ambient and personal exposure are roughly proportional. In other words exposure to background pollution is considered. Moreover the latter authors consider spatio temporal modelling to estimate the exposure burden at the country level while the former authors consider a purely spatial model at continental level.

In this paper we go on in this direction by extending the population exposure distribution introduced by Finazzi et al (2011) to the trans-Gaussian mutivariate case and by proposing a new method for computing its estimation uncertainty using methods that can handle large datasets. In particular the trans-Gaussian approach used is an evolution of Rister and Lahiri (2013) which proposes a bias correction based on Bootstrap. In particular, considering data for the European continent, we estimate multipollutant concentrations using a robust semiparamtetric extension of the model introduced by Fassó and Finazzi (2011) and Finazzi and Fassò (2014), which has been proven to successfully handle continental size airquality datasets. Model outputs are than used extensively to compute the exposure distribution as a tool for summarizing the risk related to air quality for European countries and some metropolitan areas.

2 Data

The European exposure data used here are daily data of airborne pollutants over Europe from background monitoring stations. In particular, we have seven pollutants for 3-years $(2009-2011)$, and we consider benzene (C_6H_6), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matters $(PM_{10}$ and $PM_{2.5})$ and sulphur dioxide (SO_2). The network is heterogeneous with extensive missing data, as shown in the in Figure 1 and Table 1, where we have more than 9×10^6 response variable observations with and more than 2×10^6 missing data. Covariates include Meteo (Wind speed, Pressure and Air Temperature), Land Use, Elevation (Alt) and Population, Weekend effects (WE).

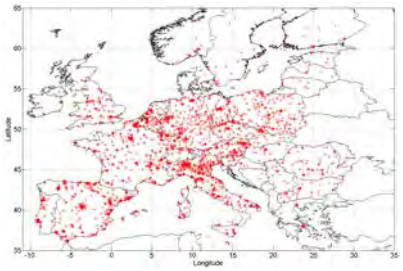


Figure 1: Monitoring network of NO_2

Subset name	Subset size	Subset use
D1	2/3	$STEM$ estimation $\rightarrow \hat{\Psi}$
D2	1/6	TG mixture estimation $\mid \hat{\Psi} \rightarrow \left(y_i \mid \hat{\xi}_i\right)$
D3	1/6	Model performance of $\hat{y}$ and $\check{y}$

Table 2: Data usage.

3 TG-STEM Model

We use a seven-variate trans-Gaussian model (TG-STEM) given by

$$\xi_i(u) = \alpha_i \omega_i(u) + \beta_{i,0}(t) + \beta_{i,1}(t)' Meteo(u) + \beta_{i,2}(t) Alt(s) + \beta_{i,3}(t) Pop(s) + \beta'_{i,4} WE(t) + \varepsilon_i(u)$$

where $u = (s,t)$, $s \in D$, for some spatial domain D, $t = 1,2,...,n$; ξ_i is the standardized log-transformation of data y_i, $i = C_6H_6, CO, NO_2, O_3, PM_{10}, PM_{2.5}$ and SO_2 in $\mu g/m^3$; $\beta_{i,j}(t) = \beta_{i,j} + Z_{ij}(t)$ are time varying coefficients; $Z = [Z_{i,j}]$ is a Vector Markovian process; $W = (\omega_1, ..., \omega_7)$ is a linear coregionalization model ε_i are independent Gaussian errors.

Model selection and fitting are performed considering the original large European dataset D discussed in Section 2 and randomly splitting in three parts as described in Table 2. In particular in D1 estimation is performed using the EM algorithm in Finazzi and Fassó (2014). While, in D2 a nonparametric back transform is developed, while D3 is used for crossvalidation.

4 Exposure distribution

In Figure 2, we compute high resolution dynamic maps using TG-STEM $\check{y}(s,t) = E\left(y(s,t) \mid \hat{\xi}(s,t)\right)$ where $\hat{\xi}(s,t) = E(\log y(s,t) \mid Y)$ and, using plug-in approach, we compute the daily exposure distribution for each pixel, uncertainty included, namely

$$\hat{F}_i(y,t) = \frac{n\left(\hat{y}_i(\mathcal{B}_r,t) \leq y,\ r = 1,...,R\right)}{n(\mathcal{R})}.$$

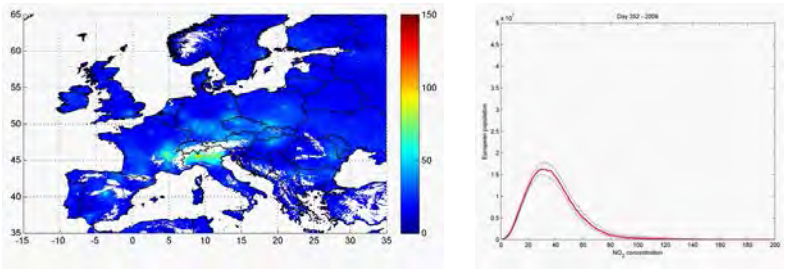


Figure 2: Mapping (left panel) and exposure distribution (right panel) for NO_2, day 352/2009.

We compute these distributions at various aggregation levels, in space from province to country, in time from day to year, and make various comparisons of European countries. For example in Figure 3, we compare three important metropolitan areas. The multivariate approach has two advantages. On the one

side allows to improve spatial information for poorly monitored pollutants such as benzene. Moreover it allows to asses joint exposure distribution which is important for health protection.

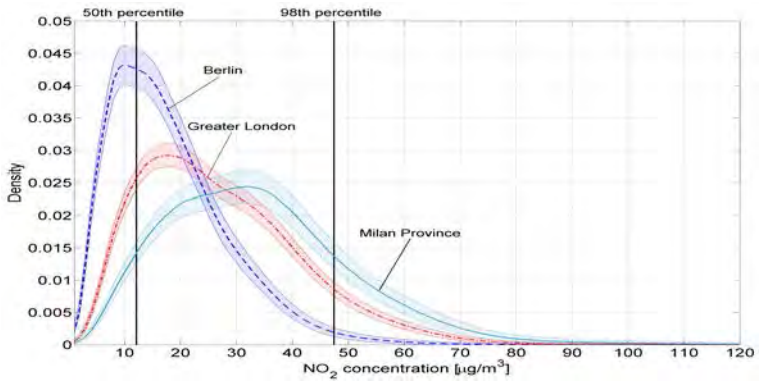


Figure 3: Exposure distribution for NO_2 in 2009, with 95% confidence bands. Vertical bars: $50th$ and $98th$ EU-percentiles.

References

[1] Fassò A., Finazzi F., (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*. Vol. 22:6, 735-748.

[2] Finazzi F., Fassò A. (2014). D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. *Journal of Statistical Software*. **62**(6), 1–25.

[3] Finazzi F, Scott M.E, Fassó A. (2013). A model based framework for air quality indices and population risk evaluation. With an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society, series C*. **62**(2), 287-308.

[4] Jahnan H.K., Kraemer A., Chen X.C., Chan C.Y., Engling G., Ward T.J. (2013). Ambient and personal PM2.5 exposure assessment in the Chinese megacity of Guangzhou. *Atmospheric Environment*, 74, 402-411.

[5] McBride S.J., Williams R.W., Creason J (2007). Bayesian hierarchical modeling of personal exposure to particulate matter. *Atmospheric Environment*, 41:29, 6143-6155.

[6] Rister K and Lahiri SN (2013). Bootstrap based Trans-Gaussian Kriging. *Statistical Modelling* 13(5&6): 509–539.

[7] Shaddick G., Yan H., Salway R., Vienneau D., Kounali D. & Briggs D. (2013). Large-scale Bayesian spatial modelling of air pollution for policy support, *Journal of Applied Statistics*, **40**(4), 777–794,

[8] Zidek, J. V., Shaddick, G., White, R., Meloche, J. and Chatfield, C. (2005). Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution. *Environmetrics*, **16**, 481–493.

Spatial bias analysis for the Weather Research and Forecasting model (WRF) over the Apulia region

F. Fedele [1,2,*], A. Pollice[3], A. Guarnieri Calò Carduccci[1], R. Bellotti[2,4]

[1] *Apulia Region Environmental Protection Agency (ARPA Puglia), Corso Trieste 27, IT-70126 Bari, Italy; f.fedele@arpa.puglia.it, a.guarnieri@arpa.puglia.it*

[2] *Dipartimento di Fisica, Università degli Studi di Bari, via Orabona, 4- IT-70125 Bari, Italy;*

[3] *Dipartimento di Scienze Economiche e Metodi Matematici, Università degli Studi di Bari, Largo Abbazia S. Scolastica, 53 - IT-70124 Bari, Italy; Alessio.Pollice@uniba.it*

[4] *Istituto Nazionale di Fisica Nucleare, Via Orabona 4, IT-70125 Bari, Italy; Roberto.Bellotti@ba.infn.it*

[*] *Corresponding author*

Abstract. *The Weather Research and Forecasting model (WRF) based on CORINE land-cover database has been used to simulate 10m wind speed and 2m temperature over the Apulia region. A validation procedure against ground data from 48 monitoring stations for both a winter and a summer period inform on the spatial distribution of the model bias. A preliminary analysis based on three indices of model performance revealed that the summer period is better simulated than the winter one. Computation of Moran's index for the spatial distribution of the model bias shows that the summer 10m wind speed bias results to be weakly autocorrelated while this autocorrelation is absent for both winter and summer 2m temperature as well as for winter 10m wind speed. On the other hand, semivariogram analysis shows a slight correlation with a range value less than 2000m for 2m temperature's winter and summer biases.*

Keywords. *Spatial Bias; Numerical Weather Prediction Model; CORINE land cover*

1 Introduction

Among Numerical Weather Prediction models a distinction is made between Global models, which usually have a coarse resolution and do not allow to simulate small scale features, and mesoscale models which cover only a small part of the planet but with a higher resolution. WRF is an example of mesoscale models [1]. Some atmospheric processes occur at spatial and temporal scales not resolved even for these high-resolution models and are represented by some physical parameterizations based on several approximations. This aspect, together with the uncertainties of the initial and boundary conditions provided by Global models, leads to the introduction of a bias in the model outputs [2, 3]. Many studies have been carried out to demonstrate the existence of a bias for 10m wind speed and 2m temperature, for several physical parameterizations. In the present study WRF model simulations for a summer and a winter period (January 2013 and July 2013) are validated against data from a set of ground stations quite
uniformly distributed over the Apulia region area. The aim of the analysis is to study the spatial pattern of the 10m wind speed and 2m temperature biases for WRF simulations.

1 Materials and methods

In the present study, simulations have been performed with the WRF model [1], developed in a cooperative effort coordinated by the National Center for Atmospheric Research (NCAR). It is a state-of-the-art numerical weather prediction system that solves the fully compressible, non-hydrostatic Euler equations forming a system of partial differential equations used to simulate the dynamic of atmospheric processes. This system is solved on different vertical and horizontal levels of the simulation domain. Below a certain vertical level, within the Planetary Boundary Layer (PBL), many processes are characterized by a space-time resolution finer than that reachable by the model and therefore these processes are described by some parametrizations. For the simulations discussed in the present paper, the model has been implemented in a one-way nesting configuration, using the MYJ parametrization of the PBL [4]. The simulation domain has 108x108 grid points and a 16 km resolution covering the central Mediterranean. The model is forced with the 3-hourly Global Forecast System (GFS) forecasts, taken as initial and boundary conditions.
The default WRF setup includes the land cover database developed and distributed from the U.S. Geological Survey (USGS) LandCover Institute (LCI) [5], which is made of 24 land cover classifications with horizontal resolution down to 1 km (last update in 2000). The model setup implemented in the present study is based on the CORINE land cover database [6] which is characterized by higher resolution and more updated categories. CORINE has been produced by the European Environmental Agency for the 28 Member States and other European countries and includes 44 land cover classifications with a resolution down to 250 meters (last update in 2006).
Simulations of the hourly data for 2m temperature and 10m wind speed during a summer (July 2013) and a winter (January 2013) period have been performed and model outcomes have been validated against ground data from a set of meteorological monitoring stations. The validation database is provided by the Agrometeorological Service of the Apulia Region (ASSOCODIPUGLIA). The total number of available stations endowed with the set of sensors we are interested in is 58. The application of standard exploratory data analysis for the year 2013 has revealed that only 48 stations show reliable trends for the selected meteorological parameters thus meeting the requirements for the further analysis.

2 Results

A preliminary evaluation of the overall model performance was done by means of root mean squared errors (RMSE's), correlation coefficients and bias standard deviations resulting from the comparison of the simulated 10m wind speed and 2m temperature databases with the ground data at each monitoring station. All these values are simultaneously reported in the Taylor diagrams of Figure 1. It is evident that the best performance of the model is obtained for July 2m (black dots) temperatures. Even if the wind speed performances are quite similar, a better performance is obtained for July (black dots) measures in terms of RMSE's for most of the stations.

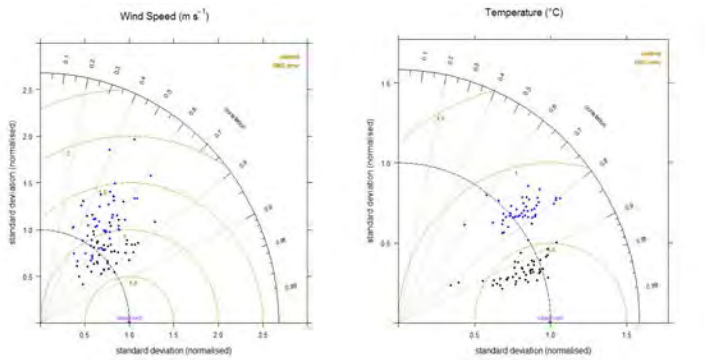


Figure 1: (left) 10m wind speed Taylor Diagram for July 2013 (black dots) and January 2013 (blue dots). (right) 2m temperature Taylor Diagram for July 2013 (black dots) and January 2013 (blue dots). Each dot represents a monitoring station.

To gain a further insight on the spatial behavior of model prediction, the mean model bias (MB) for each station is calculated and its spatial distribution has been analyzed. At the *i*-th station MB is defined as $\mathrm{MB}_i = \frac{1}{T}\sum_{t=1}^{T}(M_{it} - O_{it})$ where O_{it} and M_{it} respectively represent the t-th observed value and the corresponding simulated value at the nearest grid point and *T* is the total number of pairs.
A first check of the WRF MB spatial distribution over the Apulia region is obtained by maps in Figure 2, including MB's for the summer and winter period and for both selected meteorological parameters. As regards 10m wind speed bias, it can be noted that the summer period is the one with lower MB's, while the winter period results to be the one with the lower MB's for 2m temperature. In the case of wind speed summer MB a poor spatial variability is observed, while the other three mean biases show a higher degree of spatial heterogeneity.
To measure the spatial autocorrelation of these biases, two approaches have been used: the visual inspection of the semivariogram and Moran's I hypothesis testing [7]. In Figure 3 the semivariogram of 2m temperature and 10m wind speed biases for both periods are reported. As can be noticed, 10m wind speed bias doesn't show evidence of spatial autocorrelation, however a weak spatial trend is present in both periods. As regards 2m temperature's bias, a slight spatial correlation is observed with a range value less than 20 Km.
Moran's index I gives an alternative measure of the global spatial autocorrelation for continuous data, based on the definition of a spatial neighborhood leading to a spatial weight matrix. Moran's index lends itself to the calculation of Z-scores and relative p-values to test for presence of spatial autocorrelation.

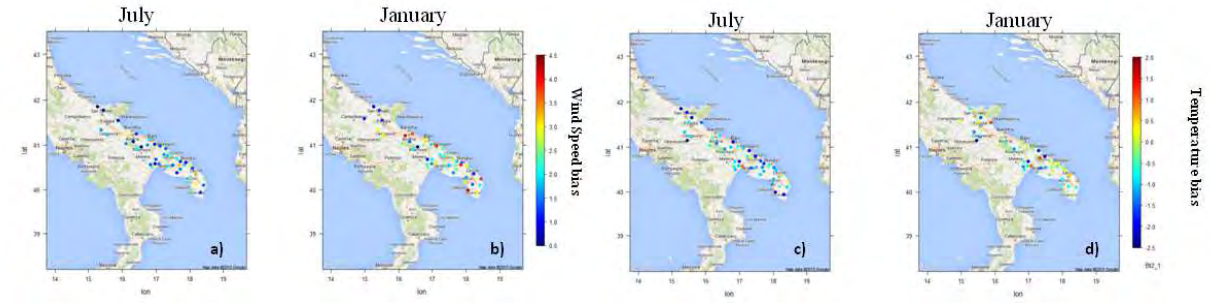


Figure 2: a) and b) 10m wind speed bias map on July 2013 and January 2013 respectively; c) and d) 2m temperature bias map on July 2013 and January 2013 respectively.

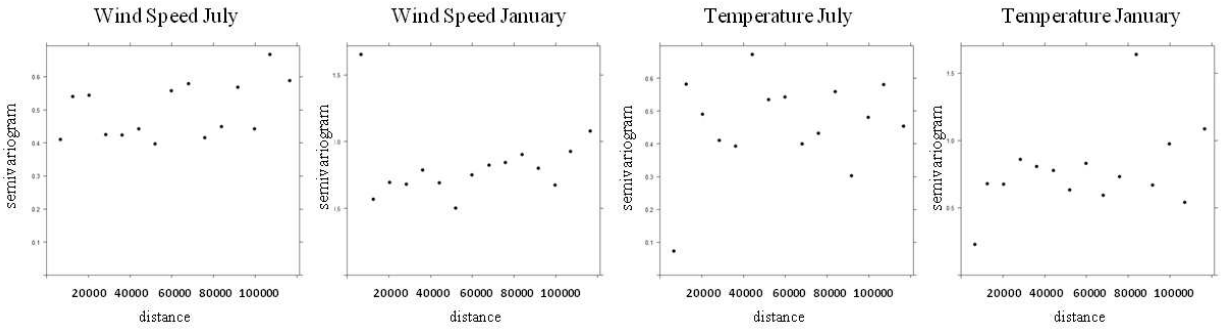


Figure 3: Semivariograms of 10m wind speed and 2m temperature biases.

Moran's I index was calculated for nine alternative definitions of the distance weight matrix obtained combining maximum and minimum distance bounds, respectively 38000 m, 45000 m, 50000 m and 6000 m, 8000 m, 10000 m. The lowest minimum distance is approximately equal to the minimum distance between stations, the lowest maximum distance is the one for which the number of neighbors is different from zero for all stations, the upper maximum distance is the distance distribution's mode.
In Figure 4 the different Moran's I values and the corresponding p-values are reported for all the weight matrixes. It is evident that the different specifications lead to similar results and the only case in which Moran's I value is significantly different from zero, though with quite a low value slightly above 0.1, is the one relative to the 10m wind speed bias for July period. This outcome implies that the spatial

patterning of this variable shows a slightly positive global spatial autocorrelation.
In conclusion, the WRF setup results to be effective giving results which are independent from the validation point as bias analysis doesn't show strong evidence of spatial autocorrelation. Moreover, as shown in Taylor diagrams, there is a better agreement between ground observations and WRF July outputs than January ones. The highest correlation is obtained for July 2m temperature.

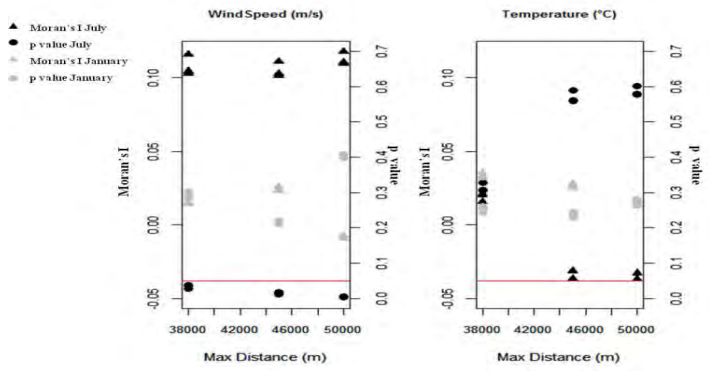


Figure 4: Moran's I and p-values for different distance matrixes for 10m wind speed bias (left) and 2m temperatures bias (right). Black marks are for July and grey marks for January. Triangles represent Moran test outcomes for each maximum distance (along x axis) and for each minimum distance and circles represent the associated p values. Red line marks the threshold of 0.05 for p value.

Acknowledgments. Authors gratefully acknowledge the Regional Association of Consortiums for the Protection of Apulia (Assocodipuglia) for making data available. The computational work has been executed on the IT resources made available by ReCaS, a project financed by the MIUR (Italian Ministry for Education, University and Re-search) in the "PON Ricerca e Competitività 2007-2013 - Azione I - Interventi di rafforzamento strutturale" PONa3_00052, Avviso 254/Ric.

References

[1] Skamarock,W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M., Huang, X.-Y., Wang, W., Powers, J.G. (2008). A description of the advanced research WRF Version 3. NCAR Technical Note NCAR/TN–475 + STR.

[2] Liu, Y., Warner, T., Wu, W., Roux, G., Cheng, W., Chen, F., Delle Monache, L., Mahoney, W., Swerdlin, S. (2009). A versatile WRF and MM5-based weather analysis and forecasting system for supporting wind energy prediction. In: 23rd WAF/19thNWPConference, AMS, Omaha, NE. 1–5 June 2009, Paper 17B.3

[3] Coleman, R.F., Drake, J.F., McAtee, M.D., Belsma, L.O. (2010). Anthropogenic moisture effects on WRF summertime surface temperature and mixing ratio forecast skill in Southern California. Weather Forecast 25:1522–1535.

[4] Janjic, Z. I. (2002). Nonsingular Implementation of the Mellor–Yamada level 2.5 scheme in the CEP Meso model. NCEP Office Note 437, 61 pp.

[5] Anderson, J.R., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976. A land use and landcover classification system for use with remote sensor data. U.S. Geological Survey Professional Paper 964.

[6] European Environmental Agency (EEA), 2000. CORINE Land Cover (1:100000), NATure/ LANd Cover information package. http://natlan.eea.eu.int.

[7] Bivand, R.S., Pebesma, E., Gómez-Rubio, V., 2013 Applied Spatial Data Analysis with R UseR! Series, Springer 2nd ed. 2013, ISBN: 978-1-4614-7617-7.

International frameworks for environmental statistics and their application to climate change related statistics

Ferruzza A.[1], Vignani D.[2,*], Tagliacozzo G.[1], Tersigni S.[1] and Tudini A.[1]

[1] *Istat Italian National Institute of Statistics; ferruzza@istat.it, tagliaco@istat.it, sttersig@istat.it, tudini@istat.it*
[2] *Istat Italian National Institute of Statistics; vignani@istat.it*
[*] *Corresponding author*

Abstract. *The last decades witnessed increasing demand of statistics and accounts in order to adequately describe environmental issues. The information required cover a wide range of interlinked statistical domains. For Climate Change (CC), linkages of environmental statistics with economic and social ones are particularly strong. Referenced frameworks - among the several national and international initiatives responding to the growing demand - have the advantage of structuring the overwhelming amount of information produced. The aim of this work is to present how two main international frameworks can be used for providing environmental statistical information, especially on CC. The first framework, the UNSD-FDES 2013 and its methodological 2014-2016 developments, is based on a multi-purpose conceptual and statistical approach, defining standardized concepts, definitions and methodologies. The second is the System of Environmental Economic Accounting 2012 - Central Framework (SEEA-CF) providing the first international statistical standard for environmental-economic accounting. In the UN-ECE context, FDES and SEEA are both primary sources in the work to define an internationally comparable set of key CC related statistics and indicators. A main objective of the UN-ECE Groups is to enhance the role of NSOs in the development of statistics on CC related phenomena. Istat provides a significant contribution to these frameworks development in building and implementing harmonized methods and definitions. The challenge is to adequately transform data into environmental statistics, relevant to official statistics production, ensuring a coherent system at national and international level, suitable to meet the increasing information demand on environment and especially on Climate Change.*

Keywords. *Environment; Climate Change; FDES- SEEA international frameworks; official statistics.*

1 Introduction

The demand for environment statistics and accounts is increasing along with continued environmental degradation and the challenges associated with better management of the environment. Environmental data are large amounts of unprocessed observations and measurements on the environment and related processes; National Statistical Offices (NSOs) or other parts of the national statistical system collect them by means of statistical surveys, or compile them from administrative records, registers, inventories, monitoring networks, remote sensing, scientific researches and environmental field studies.

By aggregating, synthesizing and structuring environmental and other data according to statistical

standards and methods, statistics are derived to describe the state and trends of the environment and the main processes affecting it. The wide range of information produced covers biophysical aspects of the environment and those aspects of the socio-economic system that directly influence and interact with the environment.

Climate Change (CC) provides a clear example of how the complexity of the issue at stake requires information encompassing a wide range of interlinked statistical domains. Statistical frameworks described in the following paragraphs have the advantage of organizing the overwhelming amount of information produced, thereby guiding its development and improving its quality. In the UN-ECE context, they are used to define an internationally comparable set of key CC related statistics and indicators.

2 FDES

Istat together with other NSOs took part in an Expert Group on the UNSD Framework for the Development of Environment Statistics (FDES, 2013) and is currently working on the related methodological manual.

FDES 2013 is a multi-purpose comprehensive and integrative conceptual and statistical framework. It is comprehensive because it facilitates data integration within environmental economic and social statistics, contributes to structuring and aggregating them into statistical series and indicators. It gives importance to geospatial information that, taking into account environmental issues (e.g., climate change, biodiversity loss, ecosystem health, natural disaster, population growth, etc.), enables integrated analyses according to different geographical units. FDES is integrative because it considers the other frameworks and systems such as the System of Environmental-Economic Accounting (SEEA), the Driving force–Pressure–State–Impact–Response (DPSIR) framework, the Millennium Development Goals (MDGs), the sustainable development indicator frameworks and the CC issues. FDES is also based on ecosystem concepts. An environment ecosystem[1] in a vital and healthy state constitutes a prerequisite to ensure an authentic well-being for all components of society.

FDES 2013 organizes environment statistics into a structure of six components: environmental conditions and quality; availability and use of environmental resources and related human activities; use of the environment as a sink for residuals and related human activities; extreme events and disasters; human settlements and environmental health; social and economic measures for the protection and management of the environment. Each component is broken down into statistical topics.

A Core Set of Environment Statistics with high priority and relevance has been identified. Besides the NSOs and environmental ministries, several other institutions are key players in the production of data used in environment statistics, adding elements of complexity to the challenge.

FDES facilitates their production in an internationally comparable manner. The UNSD expert group is currently developing detailed methodological guidance for the Core Set of Environment Statistics, including classifications, definitions, data collection and compilation methods. The sections on Water resources statistics, Waste statistics, Mineral resources, Energy resources, Expenditures on Environmental Protection and Resource Management, Land cover and Land Use, Biodiversity, Natural extreme events and disasters and will be available in the first months of 2016.

[1] The presence of a pristine environmental is the only durable insurance of having unpolluted water, clean air uncontaminated soils and food. These factors are also tightly linked to a sustainable energy consumption and transportation, smart cities and high quality human settlements.

3 The System of Environmental-Economic Accounting (SEEA) Central Framework

The SEEA Central Framework is the international statistical standard describing the interactions between the economy and the environment by means of three main types of accounts:

- "Physical flow accounts" (PFA), recording the supply of resources - e.g. minerals, timber, fish - from the environment to the economy, the flows of products within the economy and the flows of residuals from the economy to the environment in the form of, for example, solid waste and air emissions;

- "Asset accounts", measuring in quantity as well as monetary units, the stock of a specific environmental asset at the beginning and at the end of the accounting period and the changes (additions and reductions) during the accounting period;

- "Environmental activity accounts and related flows", concerning the monetary transactions between economic units whose primary purpose is environmental protection and preservation.

There is no specific account for CC within the SEEA-CF but, rather, all three types of accounts can be used to analyse several CC related issues, for example: Green House Gas (GHG) emissions caused by economic activities and households, in the case of PFA, water asset accounts describing the changes in precipitation regimes and their implications for water stocks in the case of asset accounts, monetary expenditure for actions and activities to reduce, prevent or eliminate GHG emissions, in the case of the accounts for monetary transactions.

The consistency of SEEA with the System of National Accounts (SNA) principles, definitions and classifications, and its comprehensive approach to the description of environmental issues such as CC make it a suitable candidate for deriving an internationally consistent and comparable set of key CC-related statistics and indicators; this is the purpose of the work of a UN-ECE Task Force, described in detail in the next paragraph.

4 The UN-ECE Task Force

With the aim of supporting the development of CC related statistics, a Task Force established by UN-ECE at the request of the CES, worked in 2012-2014 to develop a document of Recommendations for improving the statistics related to CC collected by national statistical systems.

The Task Force analysed existing reference frameworks to delineate the statistical subject areas related to CC: in addition to the already mentioned FDES and SEEA, also the DPSIR, the Natural capital approach and the IMA (Impact, mitigation and adaptation) have been reviewed.

The recommendations are grouped by three main areas: 1) data needed for GHG inventories (Emissions and Drivers); 2) data needed for other CC related statistics (Impacts, Mitigation, Adaptation); and 3) statistical infrastructure required for this work.

Regarding the first area, it is recommended to NSOs to be more aware of how the data of national statistical systems are or could be used in GHG inventories and to improve data and quality of required statistics. For the second area main recommendations are to: facilitate access to data that already exists; improve and support geo-spatial analysis, improve linking between socioeconomic and environmental data and develop new statistics based on a review of key data needs. For the third area the need of reviewing existing classification systems, registers, definitions, products is expressed.

As a follow up to the recommendations, the UN-ECE established in 2014 two groups that are currently working in parallel:

- a Steering Group having the mandate to provide direction on countries' progress in implementing the Recommendations, to identify areas that require further methodological work or where practical guidance would need to be developed, to promote sharing of ideas and good practice, for instance through the expert meetings;

- a new Task Force having the mandate to work on a set of key climate change -related statistics using SEEA and other frameworks such as FDES. The objective is to provide a key set of climate change -related indicators organized according to the five areas defined by the previous Task Force: (i) Emissions: GHG emissions and their human causes (ii) Drivers: human causes of climate change that deal with sources of emissions (iii) Impacts: impacts of climate change on human and natural systems (iv) Mitigation: efforts of humans to avoid the consequences (v) Adaptation: efforts to adapt to these consequences.

5 Conclusions

Human wellbeing depends on the environment and it is crucial to have statistical information on correlated theme such as climate change, biodiversity loss and natural resource management. Being interdisciplinary by nature, environment statistics are produced by a variety of data collecting institutions, and similarly numerous methods are applied in their compilation. The challenge is to build national capacities to adequately transform environmental data into environmental statistics within official systems and regular programs of work. Istat works hard both on the development of statistical frameworks and on their implementation in a continuous process aiming at enhancing the quality of official environmental statistics.

References

[1] UNECE, (2014). Conference of European Statisticians Recommendations on climate change-related statistics. *United Nations Economic Commission for Europe*. Geneva (CH). (http://www.unece.org/statistics/about-us/statstos/task-force-on-climate-change-related-statistics.html).

[2] FDES, (2013). Framework for the Development of Environment Statistics. *United Nations Statistics Division*. New York (USA). (http://unstats.un.org/unsd/environment/fdes.htm).

[3] SEEA, (2014). System of Environmental-Economic Accounting 2012 - Central Framework. United Nations New York (USA). (http://unstats.un.org/unsd/envaccounting/seeaRev/SEEA_CF_Final_en.pdf).

Real-time detection of earthquakes through a smartphone-based sensor network

Francesco Finazzi[1,*] and Alessandro Fassò[2]

[1] *Department of Management, Economics and Quantitative Methods, University of Bergamo; via dei Caniana n.2, 24127 Bergamo, Italy; francesco.finazzi@unibg.it*
[2] *Department of Management, Information and Production Engineering, University of Bergamo, viale Marconi n.5, 24044 Dalmine (BG), Italy; alessandro.fasso@unibg.it*
[*] *Corresponding author*

Abstract. *The Earthquake Network project implements a world-wide smartphone-based sensor network for the detection of earthquakes. The accelerometric sensor onboard each smartphone is used to detect vibrations which are immediately reported to a server. The server analyses the information coming from the entire network and when a quake is detected it is notified to all smartphone users in quasi real-time. In this work we propose and compare two solutions to the detection problem. One solution is based on a likelihood approach and the other is based on filtering.*

Keywords. *Dynamic networks; Real time monitoring; Android; False alarms; Poisson process*

1 Introduction

The Earthquake Network project (`http://www.earthquakenetwork.it/`) implements a world-wide network of smartphones for real-time detection of earthquakes. Smartphone accelerometric sensors detect vibrations which are possibly related to a quake (D'Alessandro and D'Anna, 2013). The data collected by all the smartphones are sent to a server which analyses them to discriminate real quakes from the "background noise". Since collection and data analysis are done in real-time, the earthquake is notified within few seconds. This should allow people living not too close to the epicentre to take measures before their area is affected. In this work, the data acquisition process and two solutions to the detection problem are discussed.

2 Data acquisition

Smartphones which take part to the Earthquake Network project run the Earthquake Network Android application (`https://play.google.com/store/apps/details?id=com.finazzi.distquake`) which collects and reports the data to the server for analysis. The application is able to understand when the smartphone is not in use and thus can be used as a network node. Moreover, the application tries to filter

out vibrations which, most likely, are not related to a quake and are induced by other sources of vibrations. A classic control chart is used to detect if the acceleration measured by the accelerometric sensor exceeds a threshold. When this happens, the event is reported to the server along with the spatial position of the smartphone. Additionally, each smartphone reports its state to the server every 30 minutes. This allows to estimate the number of smartphones that are enabled to detect vibrations and thus a possible quake.

3 Earthquake detection

Although the Android application filters some of the vibrations not induced by a quake, many of them are not discriminable and they are reported to the server. Thus, even when quakes are not occurring, the server constantly receives vibration events from smartphones all over the world at random times. When a quake strikes, however, it usually affects a relatively large area and then a given number of smartphones at the same time. The idea is to detect a quake when, for a given area and at a given time, the instant rate of the vibration events exceeds a threshold.

In order to simplify the discussion, we consider here a fixed spatial area (e.g. a city or a small region) and we assume that, when a quake occurs, the entire area is affected. The arrival times of the vibration events are assumed here to be a Poisson process. It follows that the detection problem can be solved either studying the number of events in a given time interval or studying the inter-arrival times of the events. This leads to two approaches which are discussed hereafter.

3.1 A likelihood approach

Let $\{N(t), t > t_0\}$ be the stochastic point process describing the arrival time of the vibration events which is assumed to be a Poisson process with conditional intensity function $\mu(t)$. To define $\mu(t)$, we consider the number of enabled smartphones at time t, namely n_t. Although this quantity is not directly observed at time t, we observe the number ν_t of enabled smartphones sending their "I am alive" signal in the interval $(t - 30\ min, t]$. Hence, n_t can be assumed to be a conditionally independent random variable with distribution parametrized by ν_t and some additional covariates denoted by x_t. In particular, we assume n_t to be conditionally Poisson distributed with expectation

$$E(n_t|\nu_t) = \nu_t \exp(\beta' x_t).$$

Using this model we aim at detecting the occurrence of a seismic event as soon as possible. To do this we observe that there is a delay in signal transmission and seismic wave displaced perception and we consider the vibration signals in the interval $I_\tau^t = (t - \tau, t]$, for some $\tau > 0$, as signals related to the same earthquake. Extending change point detection techniques which are tailored for permanent changes and asymptotic theory, we consider here $N_\tau^t = N(I_\tau^t)$ signals and a likelihood approach based on the generalized likelihood ratio (GLR) statistic. In order to develop the above mentioned likelihood detector, the log-likelihood of the signals in the interval I_τ^t is introduced

$$\log L(\mu|t,\tau) = \sum_{t_j \in I_\tau^t} \log \mu(t_j) - \mu\left(I_\tau^t\right).$$

where t_j are the arrival times of the events in I_τ^t. Now suppose that, in absence of earthquakes, the process intensity is $\mu^0(t)$ while under a seismic event the process intensity is

$$\mu(t) = \mu^0(t) + \frac{\lambda}{\tau}$$

with $\lambda > 0$ for $t \in I_\tau^t$ and $\lambda = 0$ otherwise. The above log-likelihood has thus the following form

$$\log L(\lambda) = \sum_{t_j \in I_\tau^t} \log\left(\mu^0(t_j) + \frac{\lambda}{\tau}\right) - \mu^0\left(I_\tau^t\right) - \lambda$$

and for a fixed τ, the GLR statistic is given by

$$GLR(\tau, t) = \sum_{t_j \in I_\tau^t} \log\left(1 + \frac{\hat{\lambda}_\tau^t}{\tau\mu^0(t_j)}\right) - \hat{\lambda}_\tau^t$$

where $\hat{\lambda}_\tau^t = \max\left(0, \arg\max_\lambda L(\lambda)\right)$. The above GLR gives an earthquake warning if

$$\sup_{\tau>0} GLR(\tau, t) > h$$

for some threshold h. In particular, since $\mu^0(t)$ depends on n_t, we replace it by its expectation $E\left(\mu^0(t)|\nu_t\right) = \alpha\nu_t \exp(\beta' x_t)$.

The second likelihood detector discussed here is based on the efficient score which is given by

$$S(\tau, t) = \frac{\partial}{\partial\lambda} \log L(\lambda)\Big|_{\lambda=0} = \sum_{t_j \in I_\tau^t} \frac{1}{\tau\mu^0(t_j)} - 1$$

and the score detector gives an earthquake warning if $\sup_{\tau>0} S(\tau, t) > h$ for some threshold h, where μ^0 is defined as above.

3.2 Filtering approach

Let $t_j > t_{j-1}$ for $j = 1, ..., n$ the first arrival times of the above $N(t)$ process with $t_0 = 0$. Moreover let $X_j = t_j - t_{j-1}$ be the time between arrivals, which, using local Poisson properties are assumed to be conditionally distributed as negative exponential random variables with mean $\mu(t_j) = E(X_j|\lambda)$, such that

$$\mu(t_j) = \lambda(t_j)^{-1}$$

where λ is given by

$$\lambda(t_j) = \alpha(t_j) + \beta(t_j) m(t_j) + e(t_j) \tag{1}$$

In (1), $e(t_j)$ is a white noise process, $m(t_j) = E(\nu_t)$ is a smooth function of time and

$$\alpha(t_j) = \alpha(t_{j-1}) + A(t_j)$$
$$\beta(t_j) = \beta(t_{j-1}) + B(t_j)$$

where A's and B's are the innovations. Note that $\alpha(t_j)$ and $\beta(t_j)$ are not observed and they are estimated using the Kalman filter (see Shumway and Stoffer, 2006). Since $X_j = t_j - t_{j-1}$ are random, the innovations are not iid and their variances are modulated according to X_j. Detection of an earthquake is done monitoring the filtered $\alpha(t_j)$ and $\beta(t_j)$ using, for instance, a control chart calibrated to have a (possibly small) false-detection rate.

4 Discussion

Both the above approaches have pro's and con's. The likelihood approach requires to observe the arrival process over the window $(t-\tau, t]$ and the detection performances are influenced by the choice of τ. If τ is too small, the earthquake can be missed due to possible delays in the transmission from the smartphones to the server. On the other hand, a large τ may produce delays in the detection and notification of the quake. The filtering approach does not require to set the above window and the Kalman update is much faster than optimizing a log-likelihood function or computing the efficient score. Nonetheless, the control chart has also an inherent delay and must be calibrated carefully. Finally, a benefit of the likelihood approach is that the likelihood function is easily extended to the case of a space-time process. This should allow to implement a detector able to detect and locate quakes considering the entire global network.

Acknowledgments. This work was partially funded by the FIRB2012 project "Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions" (RBFR12URQJ).

References

[1] D'Alessandro, A. and D'Anna, G. (2013). Suitability of Low-Cost Three-Axis MEMS Accelerometers in strong- motion seismology: tests on the LIS331DLH (iPhone) accelerometer. *Bulletin of the Seismological Society of America*. **103**(5).

[2] Shumway, R. and D. Stoffer (2006). *Time series analysis and its applications, with R examples*. Springer. New York.

Collocation uncertainty in climate monitoring

M. Franco-Villoria[1,*], R. Ignaccolo[1], A. Fassò[2], F. Madonna[3] and B.B. Demoz[4]

[1] *Department of Economics and Statistics, University of Torino, Italy; maria.francovilloria@unito.it, rosaria.ignaccolo@unito.it*
[2] *Department of Management, Information and Production Engineering, University of Bergamo, Italy; alessandro.fasso@unibg.it*
[3] *CNR-IMAA, Tito Scalo, PZ, Italy; madonna@imaa.cnr.it*
[4] *Department of Physics and Astronomy, Howard University, Washington, DC, USA; bbdemoz@Howard.edu*
[*] *Corresponding author*

Abstract. *Understanding collocation mismatch is particularly relevant for atmospheric profiles obtained by radiosondes, as the balloons containing the measuring instruments tend to drift uncontrollably from their initial launch position. We propose a heteroskedastic functional regression model capable of explaining the relationship between collocation uncertainty and a set of environmental factors, height and distance between imperfectly collocated trajectories. Along this line, a five-fold decomposition of the total collocation uncertainty is proposed, giving both a profile budget and an integrated column budget. Considering the profiles as three-dimensional trajectories, we extend the model to include a trivariate smooth function that accounts for time and space mismatch. Results from a case study where we model collocation error of relative humidity and atmospheric pressure show that model fitting is improved once heteroskedasticity is taken into account.*

Keywords. *Functional linear model; Heteroskedasticity; Generalized additive models; Mixed models; Uncertainty budget*

1 Introduction

Uncertainty of atmospheric thermodynamic variables is a key factor in assessing uncertainty of global climate change estimates given by numerical prediction models [4]. Data, e.g. atmospheric pressure, temperature or water vapour, are gathered by high technology remote instruments such as radiosondes. An important source of uncertainty is related to the collocation mismatch in space and time among different observations; i.e. the difference between the measurements obtained from two instruments (at two nearby places) that are meant to measure the same environmental variable. It is important then to understand collocation mismatch and how this may depend on potential covariates. Data, recorded at different values of height as the radiosonde goes up into the atmosphere, can be considered as functional observations. This kind of functional data can be modelled as functions depending only on height. Alternatively, they can be considered as three-dimensional (3D) trajectories, as the radiosonde balloons

drift away in the atmosphere resulting in not necessarily vertical profiles. On the other hand, little reference is made in the literature to heteroskedasticity in a functional data context; the latter is important as it allows adjusting mean estimates for non-constant variability, on top of the fact that modelling the variance function itself is of interest to understand which covariates significantly affect it.

2 Methods

The work presented at the conference is based on the papers by Fassò et al. (2014) [1] and Ignaccolo et al. (2015) [2]. In the first part, an heteroskedastic functional regression approach is proposed to model the vertical profiles of collocation uncertainty for a climate variable, in relation to environmental factors, altitude of measurement and distance between trajectories. The error variance is not assumed to be constant but a function of some variables. To better understand random and systematic uncertainty (i.e. the uncertainty budget), a five-fold decomposition of the total collocation uncertainty is proposed, namely constant bias, reducible and irreducible environmental errors, sampling error and measurement error. This allows to obtain both a profile budget and an integrated column budget. In the second part, a "point based" formulation of the heteroskedastic functional regression model mentioned above is proposed. This model includes a trivariate smooth function to account for time and space mismatch, along with potential covariates. Functional coefficients of both the conditional mean and variance are estimated by reformulating the model as a standard generalized additive model and subsequently as a mixed model. This reformulation leads to a double mixed model whose parameters are fitted using an iterative algorithm (following [3]) that allows to adjust for heteroskedasticity. As a result, covariates estimates can be adjusted for non-constant variability and estimation of the functional mean is improved. Simultaneously the conditional variance is explicitly modelled, allowing to identify significant covariates.

3 Case study

The dataset consists of 32 pairs of radiosounding profiles of atmospheric thermodynamic variables measured at two locations: the Howard University research site in Beltsville, Meryland, USA (39.054°, -76.877°, 88 m a.s.l.), which is also a GRUAN site (GCOS Reference Upper-Air Network, see *www.gruan.org* and [4]), and the U.S. National Weather Service operational site at Sterling, Virginia, USA (38.98°, -77.47°, 53 m a.s.l.). The two sites are sufficiently close (52 km line distance) to consider collocation mismatch between them, and represent a similar climate regime. Data were converted to functional observations using penalized cubic B-splines with knots regularly spaced every 50 m and penalty parameter $\lambda = 1$. An illustration of the data can be seen in Figure 1.

On one hand, we focus on collocation of relative humidity and explain its profile uncertainty using time, altitude, coordinates, and the following environmental factors: water vapour mixing ratio, pressure, temperature and wind vector. This particular variable is of interest because humidity is known to have large forecast errors, even on small time and space scales. In this case, 85% of the total collocation uncertainty is ascribed to reducible environmental error, 11% to irreducible environmental error, 3.4% to adjustable bias, 0.1% to sampling error and 0.2% to measurement error. Moreover, the collocation error has an adjustable constant bias amounting to 3.4% of the total collocation uncertainty. The model performs better below 3000m of altitude and, globally, it misses only 11.4% of the collocation uncertainty for relative humidity.

The model used for relative humidity [1] considers profiles as being vertical, i.e. as functions of height only. Instead, for modelling atmospheric pressure, we use a point based version of the heteroskedas-

tic functional regression model [2], where profiles are considered as functions with three-dimensional domain (longitude, latitude and height). In this case, we model collocation error of atmospheric pressure in terms of space (longitude, latitude), time mismatch (calendar time, flight duration difference) and a number of meteorological covariates: temperature, relative humidity, water vapour mixing ratio and orthogonal wind components from both collocated radiosondes. Results show that model fitting is improved once heteroskedasticity is taken into account; the 95% confidence bands for the estimated functional coefficients become generally narrower and the functional coefficients associated to meteorogical covariates change in shape and magnitude. AIC criteria indicates that a model with trivariate functions that take into account the interaction among longitude, latitude and height and as well among distance and height is preferred to a model where all the components act additively.

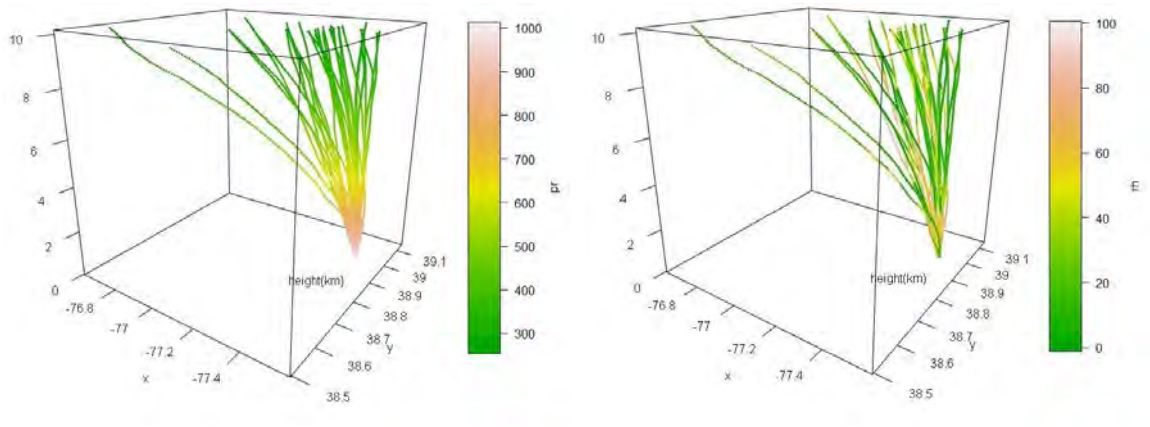


Figure 1: 3D pressure (left) and relative humidity (right) atmospheric profiles. Each curve represents a different launch at the Sterling site.

4 Conclusions

This paper presents general statistical methodology for modelling collocation uncertainty of atmospheric thermodynamic variables as introduced in Fassò et al. (2014) [1] and Ignaccolo et al. (2015) [2]. The functional regression model proposed for vertical profiles takes into account heteroskedasticity and allows the decomposition of total uncertainty budget up to five different components, namely constant bias, reducible and irreducible environmental errors, sampling error and measurement error. Moreover, the conditional uncertainty may be computed for any set of environmental conditions, providing, inter alia, more information about the factors determining the collocation uncertainty. The proposed method is self-assessing, in the sense that it is able to consider the information content of the data for the model and evaluate the size of the sampling error with respect to the other uncertainty components. In the case study considered, the collocation drift for relative humidity was found to be strongly dependent on the direction of air mass advection and not on the distance between the paired trajectories. It can be concluded that the collocation uncertainty of relative humidity is related to physical quantities and, in principle, could be reduced by inclusion of auxiliary information. The extended point based model considers the profiles as functions with three-dimensional domain, describing both conditional mean and variance as a sum of a 3D functional term and some unidimensional functional regression components. This results in

great flexibility as seen in the application to collocation uncertainty of atmospheric termodynamic profiles. The reformulation of the model as a double mixed model, with the implementation of an iterative algorithm, allows to handle the impact of covariates on conditional uncertainty by means of functional heteroskedasticity. The new 3D component is shown to improve model fitting with respect to the purely undimensional model previously considered by Fassò et al. (2014) [1] when modelling collocation mismatch of atmospheric pressure. The resulting model includes a number of terms that take into account time and space for the two collocated measurements. These effects are not linear but they smoothly change in shape along vertical direction and horizontal distance. In addition, the small unexplained collocation uncertainty changes in magnitude as explained by the heteroskedastic 3D component.

References

[1] Fassò, A., Ignaccolo, R., Madonna, F., Demoz, B. & Franco-Villoria, M. (2014). Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles. *Atmospheric Measurement Techniques* **7**, 1803–1816.

[2] Ignaccolo, R., Franco-Villoria, M. & Fassò, A. (2015). Modelling collocation uncertainty of 3D atmospheric profiles. *Stochastic Environmental Research and Risk Assessment* **29(2)**, 417–429.

[3] Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

[4] Thorne, P. W., V*ö*mel, H., Bodeker, G. et al. (2013). GCOS reference upper air network (GRUAN): Steps towards assuring future climate records. AIP Conference Proceedings 1552: 1042–1047.
doi: http://dx.doi.org/10.1063/1.4821421

Rarefaction and extrapolation with Hill numbers: a study of diversity in the Ross Sea

C. Ghiglione[1], C. Carota[2], C. R. Nava[2,*], I. Soldani[3] and S. Schiaparelli[1],

[1] *DiSTAV, University of Genova and Italian National Anctartic Museum (section of Genova); claudio.ghiglione@riftia.eu, stefano.schiaparelli@unige.it*
[2] *Department of Economics and Statistics "Cognetti de Martiis", University of Torino; cinzia.carota@unito.it, consuelorubina.nava@unito.it*
[3] *aizoOn Technology Consulting; irene.soldani@aizoon.it*
[*] *Corresponding author*

Abstract. *The Ross Sea can be considered, in a biological sense, one of the better-known areas in Antarctica due to the high number of expeditions engaged since 1899. Hundreds of mollusc species have been collected and classified along years in a unique database which is now available for study. The possibility to access such impressive information offers the opportunity to apply important results in the study of biodiversity for that area. Recent influential scientific contributions induce us to study species diversity by means of accumulation curves based on Hill numbers, i.e. the effective number of equally frequent species.*

Keywords. *Accumulation curves; Biodiversity; Extrapolation; Hill numbers; Rarefaction.*

1 Introduction

The construction of a complex and unique database, containing information related to the species richness in the Ross Sea (Antarctica) since 1899, is the result of the intensive work of an international team of ecologists. The information collected, standardized and stored along years with respect to the Phylum Mollusca in the Ross Sea, creates the biological context in which we perform the statistical biodiversity analysis described below.

Before the illustration of Hill numbers and diversity accumulation curves, a brief description of available data and their finding techniques is provided. A partial but meaningful difference, regarding the way and the aim that move researchers during expeditions, arises. Indeed, 2004 can be considered the turning point in this respect. Expeditions prior to this date were primarily focused on the realization of species inventories, without taking into account species abundances and without recording zeros for stations where any species was found. Even if this approach positively changed with the new century, these aspects generates some limitations. For instance the variable identifying species richness, i.e. the number of species found in a given sample, has only positive integer values.

Antarctica is a key location to monitor trends in biodiversity. This is a critical issue under a global warming scenario which likely would have a major impact in polar areas by introducing species from

warmer latitudes and by extinguishing stenothermal ones. Geographical shifts in species distribution patterns and temporal trends could be recognized and studied in data sets as the one here considered.

However, as already mentioned, the treatment of this type of data is not trivial. In ecology, moreover, the measurement of biodiversity is itself a complex issue deeply discussed in the literature.
Many shortcomings in the quantification of biodiversity can be defused by resorting to Hill numbers, first introduced by [3]. For a complete review over the advantages of this approach and a unification of the most important related results see [1], where main techniques for sample rarefaction and extrapolation are discussed. Here we apply some of the methods described in [1, 2] for sample-based incidence data. We draw a picture of the whole set of collected data accordingly to different gears: grab, towed and Rauschert, a specific type of dredge having a fine mesh size.

2 Methods

Traditionally, Hill numbers have been used for individual-based abundance data. Here we apply to our sample-based incidence data the methods presented in [1, 2], where a comprehensive statistical framework for the analysis of biodiversity data is provided. Therefore, our main results consist of unified diversity accumulation curves (Figures 1.a, 1.b, 1.c and 1.d). The latter are based on empirical estimates of the principal Hill numbers extended in order to incorporate information on the incidence probabilities. All these concepts are made clear in the next paragraphs.

Our data consist of a species-by-sampling-unit incidence matrix (W_{ij}) with S rows (S denotes the total number of species present in the assemblage) and $T = 456$ columns (the number of independent sampling units, i.e. discrete sampling events). Entries of the incidence matrix record the presence or absence of each species within each sampling unit: $W_{ij} = 1$ if species i is detected in the sampling unit j, $W_{ij} = 0$ otherwise. The row sum of the incidence matrix, $Y_i = \sum_{j=1}^{T} W_{ij}$, denotes the incidence-based frequency of species i, for $i = 1,..,S$. The frequencies Y_i represent the incidence reference sample to be rarefied or extrapolated in the diversity accumulation curves detailed in Figures 1.a, 1.b and 1.c. Species non detected in any sampling unit but present in the assemblage yield $Y_i = 0$. Only species with $Y_i > 0$ contribute to the total number of species observed in the reference sample denoted by S_{obs}.

Under the assumption that each species i has its own unique incidence probability π_i and that π_i is constant for any randomly selected sampling unit, each element of the incidence matrix can be viewed as a Bernoulli random variable with probability π_i that $W_{ij} = 1$ and probability $1 - \pi_i$ that $W_{ij} = 0$. This implies a Bernoulli product model for the incidence matrix,

$$P(W_{ij} = w_{ij} | \forall i = 1,2,\ldots,S,\ j = 1,2,\ldots,T) = \prod_{j=1}^{T}\prod_{i=1}^{S} \pi_i^{w_{ij}}(1-\pi_i)^{1-w_{ij}} = \prod_{i=1}^{S} \pi_i^{y_i}(1-\pi_i)^{T-y_i}.$$

Note that the likelihood under this model for W_{ij} is proportional to the likelihood under a Binomial model for $Y_1,\ldots,Y_S$. Note also that the sum of the incidence probabilities, $\sum_{i=1}^{S} \pi_i$, may be greater than 1. In [1] each parameter π_i is normalized (divided by the sum $\sum_{i=1}^{S} \pi_i$), to obtain the *relative incidence* of each species i in the assemblage; then, under the Bernoulli product model specified above, the Hill number of order q is defined as

$${}^{q}\Delta = \left(\sum_{i=1}^{S} \left[\frac{\pi_i}{\sum_{j=1}^{S} \pi_i} \right]^{q} \right)^{\frac{1}{(1-q)}}, \qquad q \geq 0,\ q \neq 1.$$

The sensitivity of ${}^{q}\Delta$ to differences in the incidence probabilities increases with q. In all cases a value ${}^{q}\delta$ of ${}^{q}\Delta$ is interpreted as the effective number of equally frequent species in the assemblage from which the sampling units are drawn. In other words, the diversity of the assemblage is the same as an ideal assemblage with ${}^{q}\delta$ species all with equal probability of incidence.

If all the incidence probabilities $(\pi_1, .., \pi_S)$ are identical, than the Hill number of all orders reduces to the species richness ${}^0\Delta$.

Hill numbers have to be regarded as theoretical or asymptotic diversities at an infinite sample size for which the true relative incidences of each i species are known. In contrast, diversity accumulation curves are based on estimates of Hill numbers of order $q = 0, 1^1, 2$, yielding the species richness, the exponential of Shannon entropy and the inverse Simpson concentration, respectively. Given the sufficient statistics $Y_0, Y_1, .., Y_S$, estimates of Hill numbers for a sample of size m are based on the incidence frequency counts $Q_k = \sum_{i=1}^{S} I(Y_i = k)$, i.e. the number of species each represented exactly $Y_i = k$ times in the incidence matrix sample $0 \leq t \leq T$.

3 Results and Conclusions

For our data we plot the main diversity accumulation curves. First, we provide the *sample-size-based rarefaction and extrapolation sampling curve* (Figure 1.a) which shows the trend of Hill numbers when the number of sampling units increases. Then the *sample completeness curve* (Figure 1.b) describes the sample completeness as a function of the sample size; it is useful to derive the sample size needed to reach a prefixed population coverage.

Finally, we show the *coverage-based rarefaction and extrapolation sampling curve* (Figure 1.c) which points out the behaviour of the species diversity as the sample coverage increases.

Finally, in order to compare the efficiency of the different gears used to collect sampling units in different expeditions, we plot the sample-size-based rarefaction and extrapolation sampling curve for grab, towed and Rauschert (Figure 1.d).

Although only eighteen sampling units are picked up with the Rauschert (due to its only recent use in the Ross Sea), which disproportionately increases the length of the 95% confidence interval (the shaded area about the curve), the superior efficiency of such a gear is apparent. Given the same sample size, the Rauschert allows to find a greater number of different species than grab and towed. Related results can be found in [4].

References

[1] Chao, A., Gotelli, N.J., Hsieh, T., Sander, E.L., Ma, K., Colwell, R.K., and Ellison, A.M. (2014). Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, **84**(1), 45–67.

[2] Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L., and Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**(1),3–21.

[3] MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews* **40**, 510–533.

[4] Schiaparelli, S., Ghiglione, C., Alvaro, M.C., Griffiths, H.J., and Linse, K. (2014). Diversity, abundance and composition in macrofaunal molluscs from the Ross sea (Antarctica): results of fine-mesh sampling along a latitudinal gradient. *Polar biology* **37**(6), 859–877.

[1] We shortly denote with 1 the limit $q \to 1$

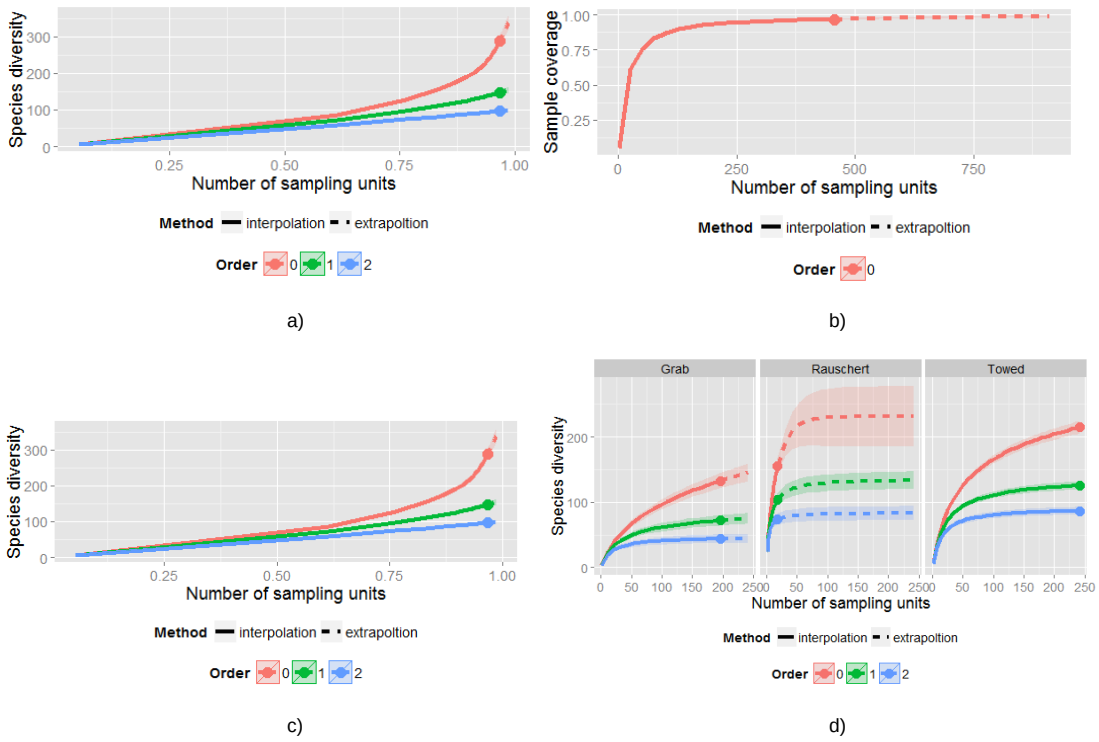


Figure 1: **a)** Sample-size-based rarefaction and extrapolation sampling curves: estimates of the number of species found in a random set of t $(t < T)$ sampling units (solid curves/rarefaction) or in an augmented set of $(T + t^*)$ $t^* > 0$ sampling units from the assemblage (dashed curves/extrapolation). **b)** Sample completeness curves. **c)** Coverage-based rarefaction and extrapolation sampling curves. **d)** Sample-size-based rarefaction and extrapolation sampling curves: grab, Rauschert and towed comparison.

Integration of different electronic nose technologies in recognition of odor sources in a solid waste composting plant

P. Giungato[1,*], P. Barbieri[2], F. Lasigna[3], G. Ventrella[1], S. C. Briguglio[2], A. Demarinis Liotile[1], E. Tamborra[1], G. de Gennaro[1]

[1] *Chemistry Department, University of Bari "Aldo Moro", Via Orabona, 4 - IT-70125 Bari, Italy; pasquale.giungato@uniba.it; gianluigi.degennaro@uniba.it; annamaria.demarinis@uniba.it; gianrocco_ventrella@yahoo.it; eliana.tamb@live.it*

[2] *Chemical and Pharmaceutical Science Department, University of Trieste, via Giorgeri 1, IT-34127 Trieste, Italy; barbierp@units.it;saracarmela.briguglio@phd.units.it*

[3] *Italcave SpA, via per Statte, 6000 - 74123 Taranto, Italy; tecnici.discarica@italcave.it*

[*] *Corresponding author*

Abstract. *Due to the continuous expansion of urban areas, the problem of emissions in the atmosphere of odors from solid industrial waste composting plants, are often cause of dissatisfaction and complaints by the communities surrounding emission sources.*

Characterization of emission sources by electronic noses is becoming a valuable approach in the management of odor emission, as are required high time resolution instrumental approaches and fast intervention on identified critical wastes, by using abatement systems.

In this paper the authors compare complementary technologies: MOSs and polymer/black carbon (Nano Composite Array – NCA) based sensors electronic noses to monitor odors emitted from an industrial solid waste composting plant, in the aim to implement integrated policies for a better management of composting operations.

10 MOS sensors in the PEN3 (Airsense), operating at high temperature and 32 polymer/black carbon (Nano Composite Array – NCA) based sensors in the Cyranose 320 (Sensigent), operating almost at ambient temperature, were tested on samples collected above three odour sources in the composting plant: biogas, sludge and urban waste.

The integrated dataset obtained from measures were explored by Principal Component Analysis and Discriminant Analysis to identify sensor discrimination capabilities, strengths and weaknesses of the technologies used.

The results obtained highlight the advantages of monitoring the composting process with a multi-tech sensor approach, in order to provide complementary information useful to better discriminate the emissions from a waste composting plant.

Keywords. *Electronic nose, MOS, polymer/black carbon, Nano Composite Array, composting plant.*

1 Introduction

Due to the continuous expansion of urban areas, the problem of emissions in the atmosphere of odors from industrial and municipal solid industrial waste composting plants, are often cause of dissatisfaction and complaints by the communities surrounding emission sources. The problem of olfactive nuisance is characterized by considerable complexity as any substances that compose the odorous mixture, produce additive, antagonistic or synergistic effect to olfactory perception [1]. Electronic noses, initially developed as instruments capable to mimic the human olfactory system, are limited by their lack of specificity (as they detect both odorous and odorless volatile compounds), lack of efficiency at remotely located sites, and remain promising instruments to monitor the transient odour level near the source so it could serve as input to mathematical dispersion models that can predict odour concentrations at remote locations together with accurate meteorological data [2,3]. In an industrial waste composting plant, the complexity of the system is enhanced by the lack in the homogeneity of the processed wastes, the numerous variables related to meteorological conditions and the particularity of the emitted odorants. Gaseous emissions in composting facilities are typically constituted by nitrogen- based compounds, sulphur-based compounds and a wide group of volatile organic compounds (VOCs) the latter emitted at the early stages of process i.e. at the tipping floors, at the shredder and during the initial forced aeration composting period [4-7]. Dimethyl sulphide (DMS), dimethyl disulphide (DMDS), limonene and α-pinene were the most significant odorous VOCs at a wastewater sludge composting facility; sulphur compounds were attributed to incomplete or insufficient aeration during composting, the terpenes to wood chips used as bulking agent [8]. Microbial activities during the aerobic decomposition of food wastes can produce peak emissions of sulfur compounds as dimethyl disulfide (DMDS), dimethyl sulfide (DMS), methyl 2-propenyl disulfide, carbonyl sulfide, methyl 1- propenyl sulfide and H_2S [9-10]. Recently new sensing technologies are being developed, as the polymer/black carbon (Nano Composite Array – NCA), to improve selectivity of sensors to specific odorous chemicals. In this case each sensor consists of conductive thin films deposited across electrodes on a ceramic substrate. When the film is exposed to a vapor-phase analyte, the polymer matrix acts like a sponge and absorbs it causing an increase in resistance. There is a lack in the scientific literature about comparative measurements of such emission sources with different sensor technologies. In this paper the authors aims to compare these two complementary technologies as Metal Oxide Semiconductors (MOSs) and polymer/black carbon (Nano Composite Array – NCA) based sensors electronic noses, to monitor odors emitted from an industrial solid waste composting plant, in the aim to recognize the emission sources and implementing integrated policies for a better management of composting operations.

2 Material and methods

The industrial waste composting plant is located in the city of Taranto, Apulia, in the south-eastern part of Italy and is operated by Italcave SpA. The electronic noses used were the PEN3 (Airsense), operating at high temperature with an array of 10 MOS sensors and the Cyranose 320 (Sensigent), operating at 42°C, with an array of 32 polymer/black carbon (Nano Composite Array – NCA) based sensors. Three sources were individuated: biogas emitted from wells disconnected to the captation network (referred as *biogas*), a waste having CER 19.12.12 (by-products of mechanical treatment of urban wastes, with no organic fraction, referred as *solid)* and the CER 19.08.05 sludge pressed and dehydrated from treatment of urban wastewater, referred as *sludge*. Samples were placed in a Nalophan 8L bag, with the lung technique, and sniffed by the two electronic noses in a randomized way. The signal of the sensors was the integral of the electrical signal (PEN3) and the relative variation of resistance DR/R_0 (Cyranose 320) during the acquisition time. Five samples of each source were collected and the integrated datasets obtained were explored by Principal Component Analysis and Discriminant Analysis, using R software package (version 3.1.2 - 2014; The R

Foundation for Statistical Computing©) together with devtools, ggbiplot and MASS libraries, in order to identify strengths and weaknesses of the sensor technology [11].

3 Results and discussion

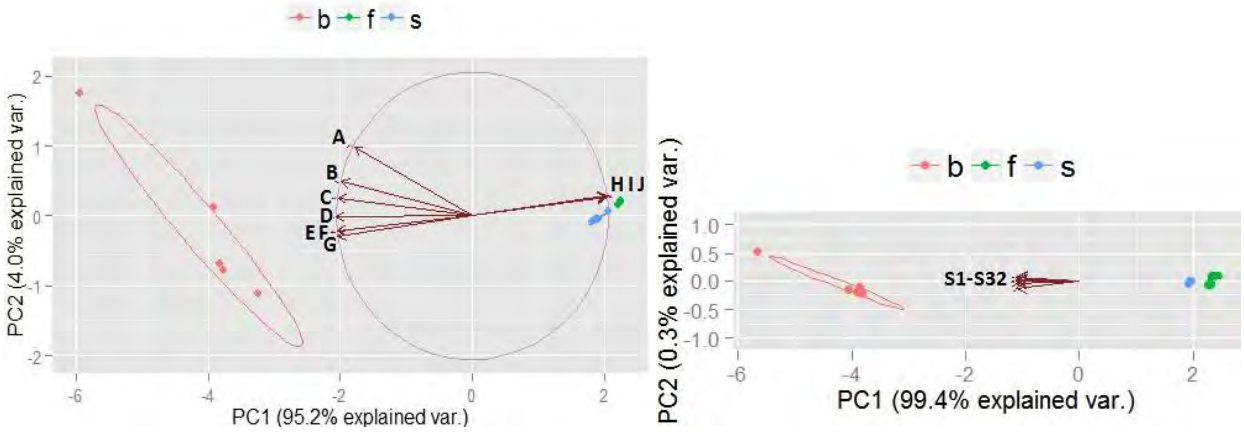


Figure 1. First two principal components of the three sources sensor array signals, using PEN3 Airsense (left) and Cyranose 320 Sensigent (right). Legend: b= *biogas*; s=*solid*; f= *sludge*; A=Broadrange, B=Sulph-chlor, C=Sulphur-organic, D=Broad-methane, E=Methan-aliphatic, F=Broad-methane, G=Hydrogen, H= Aromatic 1, I= Aromatic 3, J= Arom-aliph.

As reported in figure 1 most of the sensors of the PEN3, especially those reported by the producer as sensible to "hydrogen", "methane" and sulphur-containing gases ("sulph-chlor" and "sulphur-organic") points towards biogas scores, whereas "Aromatic 1", "Aromatic 3" and "Arom-aliph" sensors points towards sludge and solid wastes scores, demonstrating higher heterogeneity in response of this MOSs sensor technology. Cyranose 320 sensors (from "S1" to "S32") are collinear and point towards biogas scores, demonstrating higher sensor selectivity response towards these emissions. In table 1 are reported results of Linear Discriminant Analysis and cross validation for the two e-nose sensor arrays, for the combination matrix that can be obtained by integrating 10+32 sensors of both e-noses and processing it as a unique array. To improve array selectivity in this specific application, instead of using too a large array of sensors, in practical terms, a selection of only 6 sensors chosen following the selectivity with the chemicals of the emitting source and its contribution to the principal component (Aromatic, Hydrogen, Broad-methane, Sulphur-organic, belonging to PEN3, S1, S2, belonging to the Cyranose 320) of both e-noses have been tested.

E-nose	Recognition	CV% (k=1)
PEN3	100	86.7
Cyranose 320	100	53.3
PEN3+ Cyranose 320	100	60.0
Selected sensors	100	93.3

Table 1: LDA recognition by modeling set and Cross Validation prediction (k=1) of the two e-noses sensor arrays (PEN3, Cyranose 320), that of the integration (PEN3+ Cyranose 320) and that of the selected six sensors of both e-noses.

4 Conclusions

Two selected commercial gas sensor arrays, with different technologies, MOSs and polymer/black carbon (Nano Composite Array – NCA) have been tested for real-time and on-site detection of

malodours in a waste composting plant. Field tests of the two gas sensor arrays have been performed to explore the possibilities of source discrimination. A comparison of the two gas sensing technologies in the electronic noses has been carried out showing the potentialities of the portable gas sensor-system Cyranose 320 in detecting odor nuisance and the discrimination capacity in recognize the origin of the odor of the PEN3 (Airsense). Both the technological approaches were suitable for waste composting plant odor measurements in order to assess the origin of odor nuisance in critical sites. The results demonstrate that arrays of selected low-cost gas sensors may be very useful for air-pollutants monitoring and odor control applications, provided the number of sensor is reduced and the correlation between them is as short as possible: for this reason a combination of both selected MOSs and polymer/black carbon sensors should be preferable, with a selection of most sensible sensor tailored for the specific application. This work represent the first attempt to discriminate such type of sources difficult to sample and consequently with few objects per groups with commercially available e-noses, but further efforts should be done in optimizing source recognition and to select the ri ght array of sensor with tailored technology suitable to the case study.

Acknowledgments. The authors would like to thank Apulia Regions and EU for the financial support to the research activities given by the "Bando Aiuti a Sostegno dei Partenariati Regionali per l'Innovazione Investiamo nel vostro futuro", FESR P.O. 2007-2013 and the Italcave SpA for the in-field support and hosting.

References

[1] Brattoli M., Cisternino E., Dambruoso P. R., de Gennaro G., Giungato P., Mazzone A., Palmisani J., Tutino M. (2013) Gas Chromatography Analysis with Olfactometric Detection (GC-O) as a Useful Methodology for Chemical Characterization of Odorous Compounds, *Sensors* **13**, 16759-16800.

[2] Romain A-C., Delva J., Nicolas J. (2008) Complementary approaches to measure environmental odours emitted by landfill areas, *Sensors and Actuators B* **131**, 18–23.

[3] Nagle H.T., Gutierrez-Osuna R., Kermani B.G., Schiffman S.S., (2003) Environmental monitoring. *Handbook of Machine Olfaction—Electronic Nose Technology*, Wiley–VCH, Weinheim.

[4] Eitzer, B.D. (1995) Emissions of volatile organic chemicals from municipal solid waste composting facilities. *Environmental Science and Technology* **29**, 896– 902.

[5] Clemens J., Cuhls C. (2003). Greenhouse gas emissions from mechanical and biological waste treatment of municipal waste. *Environmental Technology* **24**, 745–754.

[6] Haug, R.T. (1993). *The Practical Handbook of Compost Engineering*. Lewis Publishers, Boca Raton, Florida, USA.

[7] Cadena E., Colón J., Sánchez A., Font X., Artola A. (2009). A methodology to determine gaseous emissions in a composting plant, *Waste Management* **29**, 2799-2807.

[8] Van Durme, G.P., McNamara, B.F., McGinley, C.M. (1992). Bench-scale removal of odor and volatile organic compounds at a composting facility. *Water Environment and Research* **64**, 19–27.

[9] Wu T., Wang X., Li D., Yi Z. (2010). Emission of volatile organic sulfur compounds (VOSCs) during aerobic decomposition of food wastes, *Atmospheric Environment* **44**, 5065-5071.

[10] Kim K.H., Pal R., Ahn J.W., Kim Y.H. (2009). Food decay and offensive odorants: a comparative analysis among three types of food. *Waste Management* **29**, 1265-1273.

[11] Penza, M., Suriano D., Cassano G., Pfister V., Amodio M., Trizio L., Brattoli M., De Gennaro G. (2014). A case-study of microsensors for landfill air-pollution monitoring applications. *Urban Climate*, in press.

Improving R and ArcGIS integration

K. Krivoruchko[1*] and D. Pavlushko[1]

[1] *Environmental Systems Research Institute, 380 New York St, Redlands, CA, USA, 92373; kkrivoruchko@esri.com, dpavlushko@esri.com*

[*] *Corresponding author*

Abstract. We discuss a new approach for integrating R with ArcGIS. Later this year Esri plans to release an open source R package that provides a solution *inside the application process* for passing data between ArcGIS and R. Using this new methodology, the researchers can easily build geoprocessing tools that wrap R scripts. This new methodology will potentially support a community of people who develop and share R-based geoprocessing tools for ArcGIS.

Keywords. *R integration; ArcGIS; Geoprocessing; Bayesian statistics; Thyroid cancer.*

1 Integrating R with ArcGIS

There are several variants of R scripts usage in ArcGIS applications, see for example [1,2]. Typically, a Python script is used for data transfer between ArcGIS and R. Executing the R script and rendering the results is performed using the ArcPy and other Python modules. Finally, the ArcGIS script tool allows the user to select the required data with which to run the tool and view the results in ArcGIS.
In this paper, we discuss a new approach for integrating R with ArcGIS. Esri plans to release R package that allows data to be passed between ArcGIS and R inside the application process. Using this new methodology, researchers can effortlessly build geoprocessing tools which wrap R scripts. The new approach that wraps R scripts will be free and open source. It works efficiently by minimizing library reloading, utilizing in-memory data access, and eliminating intermediate scratch files.

1.1 How the new approach works

The R script below shows how existing R code can be integrated into a geoprocessing tool. At the beginning of the script, the user initializes the *arcgisbinding* library, which can read and convert (potentially any) GIS data into an R data frame and, optionally, into the spatial data object. Then actual R script is running and its output is added to the existing data frame or a new data frame is created. Finally, the result of calculations is saved to ArcGIS dataset for further use in the geoprocessing.

```
#### The required libraries ####
library(arcgisbinding)
#### Read GIS Data Features ####
inputFC <- "C:\\Demo\\Some_Polygons.shp"
info <- arc.open(inputFC)
#### Create Data.Frame ####
df <- arc.select(info, c("FID", "CancerCases", "Population1985"))
#### Spatial Data Object####
spObject <- arc.data2sp(df)
#### Plot Spatial Data ####
```

```
spplot(spObject)
#### begin some R script, which calculates a new variable ####
####
#### end of some R script  ####
#### Add New Column/Field to the Data.Frame ####
df$new_field <- new_variable
#### Export Spatial Data to Feature Class ####
outputFC <- "C:\\Demo\\NewResult.shp"
arc.write(outputFC, df)
```

The procedure is so simple that a new geoprocessing tool can be created in a few minutes, providing that the R script was carefully tested and it works properly. Note that if the R script uses one of the plotting commands, the R pop up window will be displayed even though R environment is not running.

1.2 How the new approach improves performance of R/ArcGIS projects

New R integration approach improves performance of R/ArcGIS projects in various ways. In the list below, we start with features, which are more important for ArcGIS users, then we explain why this integration can be useful for R users and developers, and, finally, we highlight features that can be important for both GIS and R users.

- It will expand accessibility of R in the GIS community.
- The ArcGIS user can use statistical models created in R environment without even knowing Python and R languages, providing that she trusts the R script owner.
- The R usage experience is similar to other ArcGIS analysis tools usage.
- The approach to authoring and publishing analytic web services is the same for tools written in Python and R.
- It honors settings of the geoprocessing analysis environment.
  - It provides support for reading/writing of all feature and table formats available in ArcGIS.
  - It does not require R script developers to know Python.
  - While R data packages can handle relatively large datasets, they do not provide support for the traditional database management tasks. The ArcGIS platform has native support for personal, workgroup and enterprise level geodatabases. The R statistician can leverage the data management capabilities of ArcGIS then use the ArcGIS/R bridge to seamlessly bring the data into R for in-depth statistical analysis.
  - Near real-time or streaming sensor data (stock markets, weather, geolocation) is a valuable sources of information for the R statistician. The ArcGIS platform enables real-time event-based data streams to be integrated as data sources.
  - It expands the number of R libraries users (typically, thousands of researchers are visiting webpages, which provide additional data analysis functionality for ArcGIS).
    - It honors selected features and table records during data analysis.
    - It handles the reprojection of data as needed.
    - It integrates naturally with ArcGIS Python scripting environment and ModelBuilder so that the R scripts can be used together with standard Python scripts and third-party Python libraries, such as NumPy (the package for scientific computing with Python), SciPy (a Python-based open-source software for mathematics, science, and engineering), and ProBT (extended Bayesian networks framework) adding great flexibility to solving complex GIS problems.
    - The ArcGIS platform provides a powerful and convenient mechanism for sharing analysis workflows and data. ArcGIS can package geoprocessing tools and the data used by the tools into a single compressed file (.gpk). All resources (models, scripts, data, layers, and files) needed to reexecute the tools are included in the package. This means consumers of the package can rerun the tools to produce the exact same results.

We expect that new methodology will help to build a community of people who develop and share R-based geoprocessing tools.

2 Example: Bayesian analysis of thyroid cancer in children in Belarus

We illustrate the R/ArcGIS integration with regression modeling of thyroid cancer in children using data collected in Belarussian districts several years after the Chernobyl accident, from 1986 to 1994. We want to investigate the relationships between cancer rates and some environmental factors. The main reason for thyroid cancer epidemic was irradiation by short-lived iodine radionuclides, but iodine measurements collected immediately after the Chernobyl accident are scarce to reconstruct its spatial distribution. Therefore, following many other researchers, we will use the following explanatory variables: average value of ^{137}Cs soil contamination in the administrative districts and the distance from the districts to the Chernobyl nuclear power plant.

Epidemiological data, the number of cancer cases in children in each Belorussian district and the number of children in 1986 (the population under risk), are described and provided in [1]. Sufficiently complete (about 15,000 samples) of ^{137}Cs values collected in Belarus are provided in [1] and we will estimate the average ^{137}Cs values in the districts using Geostatistical Analyst's conditional Gaussian simulation (see description of the model in [1], chapter 10). We will use the average ^{137}Cs values as the exposure variable.

There are many ways to calculate distances between the administrative districts and the Chernobyl NPP. In this exercise, we use the following algorithm:

- Estimate children population density using Geostatistical Analyst's Areal Interpolation [3].
- Sample about 1000 points from that density using Spatially Balanced Design GP Tool [4].
- Use median distance between the sampled points inside each district and the Chernobyl location.

Figure 1 shows how the algorithm above works in a ModelBuilder, an ArcGIS application used to create, edit, and manage geoprocessing models. The model uses a spatial join to add district names to the sample points; then calculates the distance from each sample points to Chernobyl; and, finally, it uses a Python script to calculate the median for all points from the same district. It produces the output table with distances for each district.

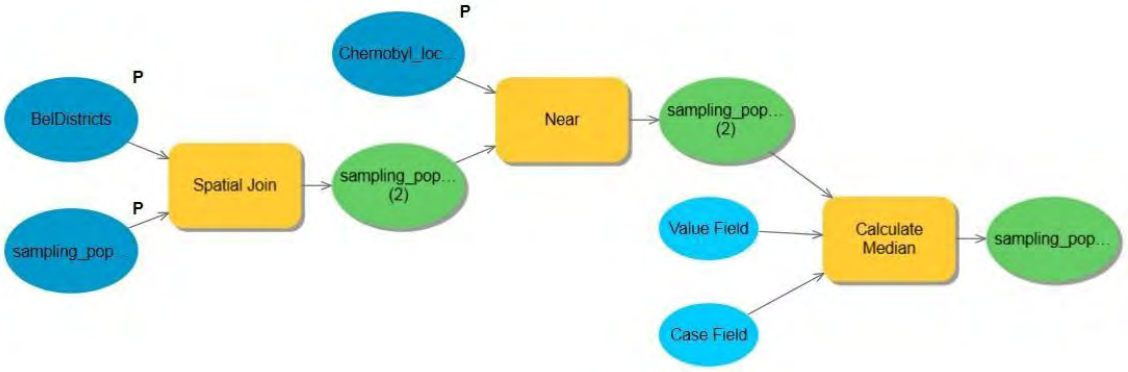


Figure 1. Distances between the administrative districts and the Chernobyl location are calculated in ArcGIS ModelBuilder.

For epidemiological regression, we use a conditionally specified prior spatial structure using a conditional autoregressive model. The model requires specification of the neighbors of each polygon and their weights. We create the neighbors list using the R *spdep* package [5].

Our model allows the regression coefficients to be spatially correlated and change locally. The R and WinBUGS [6] scripts are discussed in details in [1] (in appendix 3). A WinBUGS model is called and the inferences are summarized using R2WinBUGS R package [7]. We will provide the updated scripts and instructions for their usage in ArcGIS geoprocessing in the full paper.

Using this model, we can map the regression coefficients to see how much each covariate influences the value of the thyroid cancer risk locally. Figure 2 shows maps of the regression coefficients for ^{137}Cs soil contamination (left) and its standard error (right). We see that influence of the ^{137}Cs soil contamination covariate is gradually decreasing towards the northwest. However, the prediction standard error is very large, and we should be careful in relating the ^{137}Cs soil contamination to thyroid cancer risk.

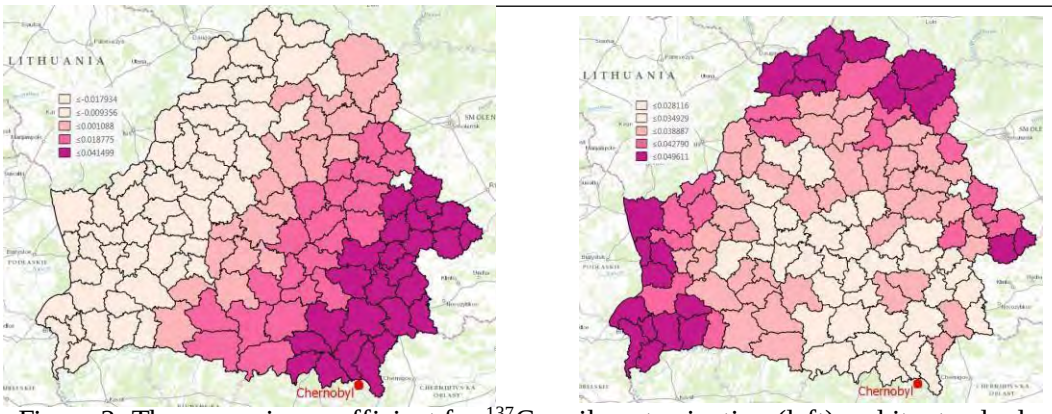


Figure 2: The regression coefficient for ^{137}Cs soil contamination (left) and its standard error (right).

Figure 3 shows the proportion of the environmental and spatially correlated components of the thyroid cancer risk for each administrative district. The geoprocessing tool, which runs the statistical model, is shown at right (note that it took 2 minutes and 17 seconds to add Bayesian statistical output to the shape file with the epidemiological data). We see that the average value of ^{137}Cs soil contamination and distance to Chernobyl do not play a significant role in the southern part of Belarus close to Chernobyl. This is because iodine ^{131}I deposition was very different from ^{137}Cs deposition since the latter radionuclide is much heavier and because the half-life of the former is very short.

It should be noted, however, that less sophisticated models found significant relationship between thyroid cancer in children and ^{137}Cs data, see [1].

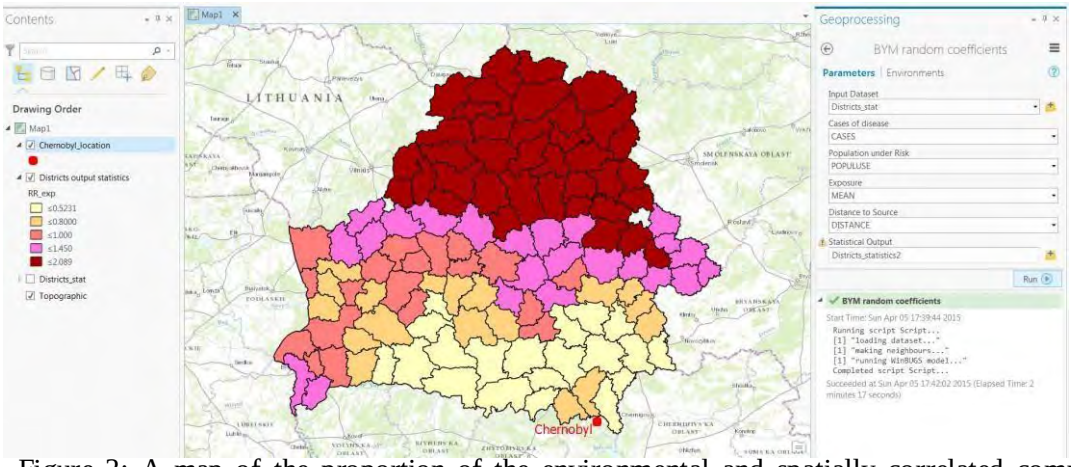


Figure 3: A map of the proportion of the environmental and spatially correlated components of the thyroid cancer risk and the geoprocessing tool, which runs the statistical model (at right.)

Note that the covariates used in this exercise are not precise and the analysis can be made more realistic by taking into account the covariates uncertainties.

References

[1] Krivoruchko K. (2011) *Spatial Statistical Data Analysis for GIS Users*. ESRI Press, 928 pp.
[2] Introduction to R scripting with ArcGIS, http://esri.ca/en/content/introduction-r-scripting-arcgis.
[3] Krivoruchko K., Gribov A. and Krause E. (2011) Multivariate Areal Interpolation for Continuous
[4] Krivoruchko K. and Butler K. (2013) Unequal Probability-Based Spatial Sampling. ArcUser Spring 2013, pp.10-17. Also available online at http://www.esri.com/esri-news/arcuser/spring-2013/unequal-probability-based-spatial-sampling
[5] Bivand R. and Piras G. (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software, 63(18), 1-36. URL http://www.jstatsoft.org/v63/i18/.
[6] The BUGS Project. http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs.
and Count Data. Procedia Environmental Sciences. Volume 3, 2011, Pages 14-19.
[7] Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software, 12(3), 1-16.

Energy-efficiency optimization of the biomass pelleting process by using statistical indicators

F. Manca[1], E. Loiacono[2], G. L. Cascella[2]*, D. Cascella[2]

[1] *Department of Education, Psychology and Communication - University of Bari Aldo Moro; fabio.manca@uniba.it*
[2] *Idea75 Srl - Via Guido de Ruggiero 1, Bari; e.loiacono@idea75.it; g.l.cascella@idea75.it; d.cascella@idea75.it*
**Corresponding author*

Abstract. *Biomass pelleting process strongly depends on a number of variables hard to be simultaneously controlled. This paper suggests a method to ensure pellets moisture optimization and process energy saving. An experimental testbed was arranged in order to validate the performance of the proposed strategy. It is based on a closed-loop control system that regulates material moisture and flow rate, but its robustness is affected by the control-loop delay (the actuator delay is about 10 minutes) and by the random arrangement of the pellets inside the cooler that strongly affects product moisture (the measurement errors are not negligible). To overcome those problems, a robust statistical approach was adopted to reach the best tradeoff between estimation accuracy and computational effort. It was derived by the well known Random Close Packing model and statistical estimator. Experimental results prove the effectiveness of the proposed approach that provides moisture errors less than 7.2% with a continuous limitation of energy consumption.*
The present work is part of Idea75's project - SEI Smart supervisor for Energy efficiency optimization of Industrial processes - funded by Regione - PO FESR 2007-2013, Asse I, Linea di Intervento 1.1. Azione 1.1.3 - Aiuti alle piccole imprese innovative di nuova costituzione.

Keywords. *Biomass pelletizing; Energy saving; Robust optimization; System identification; Closed-loop control system.*

1 Introduction

Focusing on economical and environmental advantages, biomass pelleting process represents a well known solution for the molding of waste organic materials addressed to be used as biofuels, compost, or animal feed depending on their composition. In order to ensure certain pellets requested characteristics with the lowest energy consumption, an accurate characterization of all factors that influence pelleting process is needed. Due to the presence of several input, output and system parameters, extrusion is a multiple input and multiple output process that needs an appropriate method for prediction of processing results[1]. In literature, the identification of factors affecting quality and energy consumption in pellets production is conducted for example by means of: dynamic modeling and steady state modeling[2]; statistical approaches for data processing, correlation and regression analysis[3]; Genetic Algorithms (GAs), Artificial Neural Networks (ANNs)[4, 5] and Response Surface Method (RSM)[6] for establishing mathematical relationships between input variables and product properties.
This paper focused on two different but highly correlated aspects in pellets production: moisture content and energy consumption controls. They are generally influenced by several dependent and independent

variables that cause their monitoring modestly accurate; this paper proposed a statistical approach to reduce errors and complications due to variables complexity, so allowing the system identification and the robust optimization of the used pelleting process.

2 Experimental testbed

Pelleting process synthetically consists of an initial phase in which waste materials are mixed together and then softened by addition of water and heat. The so formed mixture is conveniently compressed, dried, and cooled in order to reach a sufficiently mechanical strength and preserve pellets quality.
The used experimental testbed was based on a closed-loop control system and mainly consisted of: screw and valve respectively for raw materials and water introduction, chamber in which they are mixed, pelletizer, cooler, sensor for motor current control, microwave moisture sensor and PLC for actuators management. Also an inverter is connected to screw motor ensuring variable rotation speed and material flow rate.
Pellets to be dried and cooled are randomly dropped into a chamber and are exposed to the microwave sensor for their moisture content measurements: values are measured on different incoming product volumes before to be moved into a storage tank. Pellets random packing density strongly affects moisture values. Moreover delays in feedback measurements generates error in process control. In order to reduce these errors firstly density-moisture dependency was estimated, then a statistical method for the reduction of delay-related errors was applied. The automated experimental testbed was firstly modeled and related output were properly manipulated by using a statistical approach.

3 Model description

Process control consisted in monitoring of product moisture and motor current, respectively by means of moisture and current sensors. Those generate the feedback for the water and raw material regulation in order to reach the desired reference. The energy efficiency of the process is estimated by the following ratio: amount of processed material divided by motor current. The higher this ratio, the higher the process energy efficiency. The used testbed is reported in Figure 1.

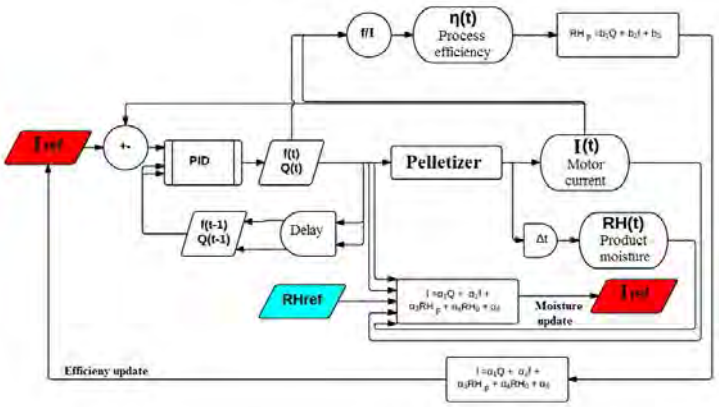


Figure 1 : Experimental testbed and related closed-loop control system.

The closed-loop control system was approximated to a linear model mainly based on current control according to relation (1).

$$I = \alpha_1 Q + \alpha_2 f + \alpha_3 RH_p + \alpha_4 RH_0 + \alpha_5 \qquad (1)$$

where I [A] is the motor current, Q [kg/s] is the raw material flow rate, f [Hz] is the frequency of the inverter connected to the screw, RH_p and RH_0 [%] respectively are pellets and raw material relative moisture contents, α_n are experimentally determined coefficients. The control system aims to keep the current as close as possible to $I_{setpoint}$. Frequency and flow rate are properly adjusted according with the feedback from final product moisture and motor current. The feedback measurements are affected by delays, consequently control errors are generated. Moisture measurements are average values between product moisture and air humidity in the empty spaces, as quantified by relation (2). Also actual product moisture is strongly affected by pellets random packing density inside the cooler. For these reasons, pellets random density and moisture-density dependency were firstly estimated.

$$RH_{sensor} = \phi RH_p + (1 - \phi) RH_{air} \qquad (2)$$

Both product density (established through the Random Close Packing model for cylindrical pellets density randomly gathered in a certain volume) and product moisture distribution are Gaussian functions. In order to reduce error on moisture measurements a statistical estimator, based on the mean value of the last n measures instead of punctual ones, was used. Variance mean value is expressed as:

$$\sigma^2 = \left[\overline{RH}\right] = \frac{\sigma^2[RH]}{(n-1)} \qquad (3)$$

Therefore increasing the number of samples, both variance and standard deviation decreased, as shown in Figure 2. Samples are the moisture contents measured on different product volumes inside the cooler; after that the product moves to the storage tank. The collection of several values is required for a more accurate estimation of actual final moisture content.
Distributions were built through Matlab fit probability distribution tools.

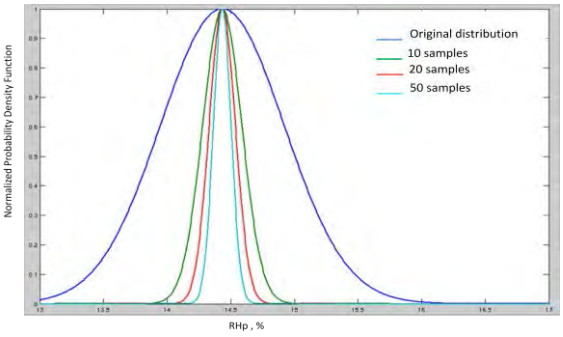


Figure 2 : Normalized error distribution plotted for different number of samples.

In order to reduce the measurement error a large number of sample is recommended; on the other hand this implies a larger delay in feedback calculation. Therefore error function on moisture measurements ξ_m decreases, error function on the process control ξ_c increases during the period between moisture calculation and action on actuators. Moreover according with Nyquist criterion for the stability determination of closed-loop systems, delay increasing over a critical value makes the loop system unstable, thus samples collection has to be properly limited.
A robust optimization technique was used to find the best number of samples for the feedback calculation. This strategy outstands results obtained by the following standard method. Both ξ_m and ξ_c depend on sample time t_s, their sum can be approximated with a continuous-time function: the best tradeoff between measure accuracy and control dynamics can be simply found by determining the

minimum of (4).

$$\frac{d}{dt_s}[\xi_m(t_s) + \xi_c(t_s)] = 0 \qquad (4)$$

4 Experimental results and conclusions

During each run of control, the use of the here presented approach guaranteed a continuous control on process variables for pellets moisture control and thus energy consumption. In fact, as shown in Figure 3, by properly regulating process variables (f and Q) into a certain convenient range, obtained moisture oscillated around the desired value RH_p = 15.2% with slight deviations only up to 7.2%.

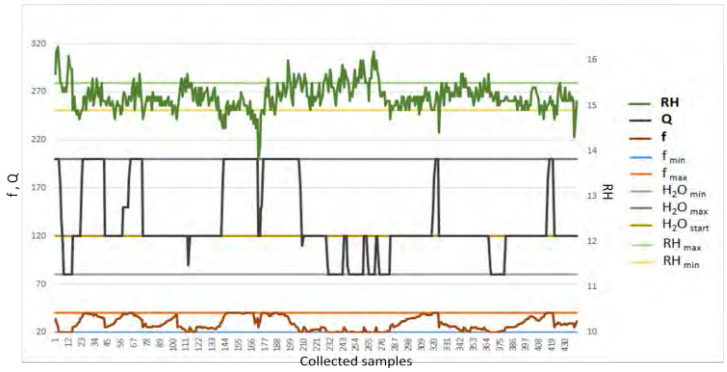


Figure 3 : Moisture trend.

This innovative statistical approach successfully faces this problem. The proposed technique minimizes errors control and guarantees more repeatability and accuracy of the process conditions. This optimized closed-loop system also ensures a convenient control of energy consumption.

References

[1] Ganjyal, G.M., Hanna M.A., Jones D.D. (2003). Modeling selected properties of extruded waxy maize cross-linked starches with neural networks. *Journal of food science*, **68**(4).

[2] Levine, L., Symes, S., Weimer, J. (2008). A simulation of the effect of formula and feed rate variations on the transient behavior of starved extrusion screws, *Biotechnology Progress* **3**, 212-220.

[3] Vigants, H., Uuemaa, P., Veidenbergs, I., Blumberga D. (2014). Cleaner pellet production – an energy consumption study using statistical analysis. *Agronomy Research* **12**(2), 633-644.

[4] Zafari A., Kianmehr M. H., Abdolahzadeh R. (2013). Modeling the effect of extrusion parameters on density of biomass pellet using artificial neural network. *International Journal of recycling of Organic waste in Agriculture* **2**(1), article No. 9.

[5] Shankar, T.J., Bandyopadhyay, S. (2004). Optimization of extrusion process variables using a genetic algorithm, *Food Bioproducts Process* **82,** (C2), 143 – 150.

[6] Shankar, T.J., Sokhansanj, S., Bandypadhyay, S., Bawa, A.S. (2010). A case study on optimization of biomass flow during single screw extrusion cooking using genetic algorithm (GA) and response surface method (RSM). *Food Bioprocess Technology* **3**, 498- 510.

Environmental SmartCities: statistical mapping of environmental risk for natural and anthropic disasters in Chile

O. Nicolis

Instituto de Estadística, Universidad de Valparaíso, Av.Gran Bretaña 1111, 234000 Valparaíso, Chile; orietta.nicolis@uv.cl

Abstract. *The main aim of the work is to build statistical environmental risk maps for natural disasters in Chile, using spatio-temporal models in order to improve the assessment, prevention and mitigation of impacts. To this end, we analyze the spatial and temporal variability of the observed points, we study the dependence from the exogenous variables, and we create risk maps. Finally, we display the results in web platforms for mobile devices. Several environmental phenomena are considered such as earthquakes, wildfires, and air pollution. In all cases the methodology is based on the assumption that data can be modeled as a spatio-temporal process, although specific models are proposed for each category.*

Keywords. *Natural disasters; Spatio temporal modeling; Hazard map.*

1 Introduction

A natural risk can be defined as the probability that a natural phenomenon result in a natural disaster, called extraordinary event. Events that can potentially result in natural disasters can be classified as earthquakes, tsunamis, forest fires, avalanches, volcanic eruptions, etc.. The probability of their occurrence may not be homogeneous in space and time. Then, spatial variations can be displayed on a map, giving rise to a environmental hazard map. Predictive maps show the probability of occurrence taking account the time of occurrence of natural events. To construct this type of maps, it is required to model risk, which can be done by considering the event occurrence as a random point process. Such kind of processes can be defined as a random collection of points falling in a specific space. In most applications, each point represents the time and/or location of an event, such as a the epicenter of an earthquake or the centroid locations of forest fires. A spatio-temporal point process is defined as a random collection of points, where each point represents the time and location of an event ([10]). Typically, the spatial locations are recorded in three spatial coordinates, such as longitude, latitude and height or depth, though sometimes only one or two spatial coordinates are available or of interest. Catalogs of spatio-temporal data may also include explanatory variables, which could may be given by a spatial function $Z(x,y)$ defined at all spatial locations (x,y) (for example, as mentioned, altitude, temperature, wind speed and wind direction in the study of wildfires) or by another spatial pattern or line segment pattern (for example, the geological faults for evaluating the earthquake risk). Any analytical spatio-temporal point process is characterized uniquely by its associated conditional rate process or conditional intensity, which is usually indicated by

(t,x,y) ([7], [4]). The conditional intensity (t,x,y) may be thought as the frequency with which events are expected to occur around a particular location (t,x,y) in spatio-temporal, conditional on the prior history, H_t, of the point process up to time t. Formally, the conditional rate (t,x,y) may be defined as a limiting conditional expectation, provided the limit exists. The behavior of a spatio-temporal point process N is typically modeled by specifying a functional form for (t,x,y), which may be estimated non-parametrically or via a parametric model (see [5], [9], and [22]). In general, (t,x,y) depends not only on (t,x,y), but also on the times and locations of preceding events. Processes that display substantial spatial heterogeneity, such as earthquake epicenters, are sometimes modeled by stationary processes in time but not space. A commonly used form for such models is a spatial-temporal generalization of the Hawkes model, known as ETAS models proposed by [14]. The conditional intensity of ETAS models can be written as:

$$\lambda_{\theta}(t,s|\mathcal{H}_t) = \mu f(s) + \sum_{t_j<t} g(t-t_j|m_j)\ell(x-x_j,y-y_j|m_j) \tag{1}$$

where the sum is over all points (t_i,x_i,y_i) with $t_i < t$. The functions μ and g represent the deterministic background rate and clustering density (with magnitude $m > m_c$), respectively. Often μ is modeled as merely a function of the spatial coordinates (x,y), and may be estimated nonparametrically as in [14]. A variety of forms has been proposed for clustering the density g ([13]; [14]; [23]). Also, different estimation algorithms have been proposed for reducing the computational time ([15]; [20]; [1]). Sometimes, the conditional intensity λ is modeled as a product of marginal conditional intensities

$$\lambda(t,x,y) = \lambda_1(t)\lambda_2(x,y),$$

where forms embody the notion that the temporal behavior of the process is independent of the spatial behavior and, in the latter case, that furthermore the behavior along each of the spatial coordinates can be seen as independent. A wide range of models describe processes in which aggregation or repulsion between events is presented, known as "shot noise" [12]. In these models, the intensity function has the form

$$\lambda(x,y) = \lambda_1(x,y)\lambda_2(t)S(x,y,t)$$

, where $\lambda_1(x,y)$ is the intensity in the space, which can be modeled as a function of environmental and climatic variables; $\lambda_2(t)$ is the temporal intensity which depends on temporal variables, and $S(x,y,t)$ is the shot noise term, which allows us to model variability.

Some environmental disasters are caused by human activities that alter normal environment. Atmospheric pollution is an example. In this case, the risk is that the concentration of a contaminant is greater than a threshold, considered dangerous to human health. Mapping the concentration of a pollutant, it is possible to identify areas most at risk than others and estimate the human exposure. The data of air pollution are usually collected by a spatial monitoring network at regular intervals (say, every hour or day or week). Thus, the data analysis has to take account temporal correlations as well as spatial correlations. Geostatistical approaches to spatio-temporal prediction in environmental science rely on appropriate correlation/covariance models ([3]). Let Z be a spatial-temporal process (i.e. concentrations of PM2.5) observed at the spatial locations $s_1,\ldots,s_n \in D$, where $s_i = (x_i,y_i)$, for $i = 1,\ldots,n$, and times $t_1,\ldots,t_m \in T$, a simple spatio-temporal model can be defined as

$$Z(s,t) = \mu(s,t) + \varepsilon(s,t) \tag{2}$$

where $\mu(s,t) = X(s,t)\beta$ is a deterministic trend component depending on the exogenous variables $X(s,t)$ (that is, temperature, relative humidity, wind speed, wind direction, land use, elevation, etc.) and $\varepsilon(s,t)$ is a zero-mean intrinsically stationary spatio-temporal stochastic process which covariance structure is normally specified by an isotropic parametric function (that is, exponential, Gaussian, Matérn). Many

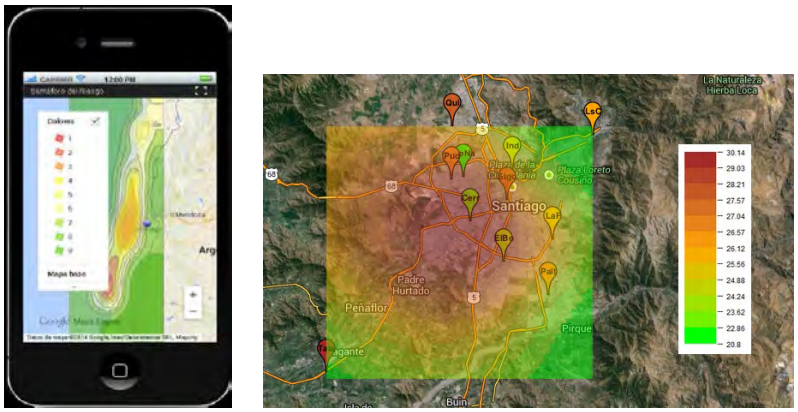
extensions of the model 2 have been proposed in the literature ([6], [18], [2]). The following model has been recently proposed by [16] for modelling the trend of air pollutant concentrations

$$\mu(s,t) = \sum_{l=1}^{L} \beta_l X_l(s,t) + \sum_{k=1}^{K} \gamma_k(s)\psi_k(t). \tag{3}$$

where $X_l(s,t)$ are spatio-temporal covariates; β_l are the coefficients for the spatio-temporal covariates; the $\{\psi_k(t)\}_{k=1}^{K}$ is a set of (smooth) temporal basis functions with $\psi_1(t) = 1$ estimated by the modified singular value decomposition (see,[8] and [21]), and the $\gamma_k(s)$ are spatially varying coefficients for the temporal functions.

2 Visualizing on web and mobil devices

The outputs of the spatio temporal models described in Section 1, can be used for visualizing the environmental hazard through web platforms for mobil devices. As an example, we show the results of the ETAS model described by Eq. 1 for visualizing the seismic risk in Chile. The map of Fig. 1 (a) represents the estimated seismicity of Chile using the earthquake catalogue from the January 2000 to May 2014. The results of the ETAS model have been classified into nine categories of colors representing the different seismic hazard rate. The GPS system of the mobile device allows to show if the user is in a high level risk position. A similar result can be obtained for assessing the daily level of air pollution as represented in Fig. 1 (b).



(a) (b)

Figure 1: (a) Estimating seismic hazard using earthquake events in the period 2000-2007 on *Google Earth* platform for mobile device. The blu circle indicate the position of the user. (b) Visualization of the estimated average PM2.5 concentrations for the month June, 2011 on *Google Map* platform.

Acknowledgments. The present work has been partially supported by Fondecyt grant 1131147.

References

[1] Chiodi, M., Adelfio, G. (2011). Forward Likelihood-based predictive approach for spatio-temporal processes. *Environmetrics*, 22, 749–757.

[2] . Clark. J.S., Gelfand A.E. (eds.) (2006). *Hierarchical Modelling for the Environmental Sciences*. Oxford University Press, Oxford, England.

[3] Cressie, N., Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, New York.

[4] Daley, D., Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*, Springer-Verlag, Berlin.

[5] Diggle, P. (1985). A kernel method for smoothing point process data, Applied Statistics 34, 138–147.

[6] Fassò A., Cameletti M., Nicolis O. (2007). Air quality monitoring using heterogeneous networks, *Environmetrics*, **18**, 245-264.

[7] Fishman, P.M., Snyder, D.L. (1976). The statistical analysis of space-time point processes, *IEEE Transactions on Information Theory*, IT-**22**, 257–274.

[8] Fuentes M., Guttorp P., Sampson P.D. (2006). Using transforms to analyze space-time processes. In B Finkenstadt, L Held, V Isham (eds.), *Statistical Methods for Spatio-Temporal Systems*, 77–150. CRC/Chapman and Hall.

[9] Guttorp, P., Thompson, M. (1990). Nonparametric estimation of intensities for sampled counting processes, *Journal of the Royal Statistical Society, Series B* **52**, 157–173.

[10] Guttorp, P.M., Brillinger, D.R., Schoenberg, F.P. (2002). Point processes, spatial. in *Encyclopedia of Environmetrics*, El-Shaarawi. A., Piegorsch, W. (eds). Wiley, New York, **3**, 1571–1573.

[11] Karr, A. (1991). *Point Processes and Their Statistical Inference*, 2nd Edition, Marcel Dekker, New York.

[12] Møller, J. (2003). Shot noise Cox processes. *Advances in Applied Probability*, **35**, 614–640.

[13] Musmeci, F., Vere-Jones, D. (1992). A spatio-temporal clustering model for historical earthquakes, *The Annals of the Institute of Statistical Mathematics*, **44**, 1–11.

[14] Ogata, Y. (1998). Spatio-temporal point process models for earthquake occurrences, *The Annals of the Institute of Statistical Mathematics*, **50**, 379–402.

[15] Ogata, Y., Zhuang, J. (2006). Spatio-temporal ETAS models and an improved extension, *Tectonophysics*, **413**, 13-23.

[16] Olives, C., Kaufman J.D., Sheppard L., Szpiro A.A., Lindström, J., Sampson P.D. (2014). Reduced-rank spatio-temporal modeling of air pollution concentrations in the multi-ethnic study of atherosclerosis and air pollution. *Annals of Applied Statistics*. **8**(4):2509–2537.

[17] Peng, R.D., Schoenberg, F.P., Woods, J. (2005). A spatio-temporal conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, **100**, 26-35.

[18] Sahu, S.K., Nicolis, O. (2008). An evaluation of European air pollution regulations for particulate matter monitored from a heterogeneous network. *Environmetrics*, **20**, 943–961.

[19] Schoenberg, F.P., Brillinger, D.R., Guttorp, P.M. (2002) Point processes, spatial-temporal. in *Encyclopedia of Environmetrics*, El-Shaarawi. A., Piegorsch, W. (eds). Wiley, New York, **3**, 1573–1577.

[20] Schoenberg, F.P. (2013) Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America*, **103**, 601-605.

[21] Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D. and Kaufman, J. D. (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, **21**, 606–631.

[22] Vere-Jones, D. (1992) Statistical methods for the description and display of earthquake catalogs, in S*tatistics in the Environmental and Earth Sciences*, A. Walden, Guttorp, P. (eds). Arnold, London, 220–246.

[23] Zhuang, J., Ogata, Y., Vere-Jones, D. (2002). Stochastic declustering of spatio-temporal earthquake occurrences. *Journal of the American Statistical Association*, **97**, 369–379.

A semi–parametric approach in the estimation of the structural risk in environmental applications

R. Pappadà[1,*], E. Perrone[2], F. Durante[3] and G. Salvadori[4]

[1] *Department of Economics, Business, Mathematics and Statistics, University of Trieste, I-34127 Trieste (Italy); rpappada@units.it*
[2] *Institut für Angewandte Statistik, Johannes Kepler Universität, A-4040 Linz (Austria); elisa.perrone@jku.at*
[3] *Faculty of Economics and Management, Free University of Bozen-Bolzano, I-39100 Bolzano (Italy); fabrizio.durante@unibz.it*
[4] *Dipartimento di Matematica e Fisica, Università del Salento, I-73100 Lecce (Italy); gianfausto.salvadori@unisalento.it*
[*] *Corresponding author*

Abstract. *In environmental applications, the estimation of the structural risk is crucial. A statistical model for the behavior of the input variables is generally required, possibly accounting for different dependence structures among such variables. Copulas represent a suitable tool for dealing with natural extremes and non-linear dependencies. Two semi-parametric procedures for the approximation of, respectively, Extreme Value and Archimedean copulas, are proposed in order to provide a model for the estimation of the structural risk. The approximating techniques are evaluated by Monte Carlo tests, and illustrated via a case study concerning a preliminary rubble mound breakwater design.*

Keywords. *Copula; Monte Carlo; Return Period; Structural risk*

1 Introduction

Recent developments in environmental statistics have shown the great potential of copulas in a multivariate risk assessment framework (see, for instance, [1, 4]). In the traditional structural approach, the response $\mathbf{Y} = (Y_1, \ldots, Y_m)$ of a given structure to some environmental (random) loads $\mathbf{X} = (X_1, \ldots, X_d)$ is evaluated via a (multi-dimensional) structural equation, generally written as

$$\mathbf{Y} = \Psi(\mathbf{X}; \boldsymbol{\theta}), \tag{1}$$

where the structure function Ψ is known, and $\boldsymbol{\theta}$ represents a set of possible covariates. Beside the knowledge of the physical response of the structure to the loads of interest, the structural approach requires a statistical model for the behavior of the variables affecting the structure, which are generally dependent. Copulas may supply valuable dependence models accounting for a wide variety of dependence patterns. A multivariate approach based on copulas is adopted, by considering two structures often used in practice: the Extreme Value (hereafter, *EV*) copulas for dealing with rare (catastrophic) events, and the Archimedean copulas for their desirable analytical properties. The selection of an appropriate

copula model still presents some troublesome issues. As a valid alternative, this work provides two semi-parametric approximation procedures to, respectively, EV and Archimedean copulas, which can be used in the estimation of the structural risk. A "minimal" approximating strategy is adopted, involving the least number of parameters and presenting the fewest fitting difficulties.
In the following, a classical problem in coastal engineering is considered to illustrate the procedures. The problem concerns the preliminary design of a rubble mound breakwater and the target is to compute, for prescribed Return Periods (hereafter, *RP*), the weight $\mathbf{Y} = W$ of a concrete cube element forming the breakwater structure, assuming that the environmental load is given by the pair of dependent variables $\mathbf{X} = (H, D)$, where H represents the significant wave height (in meters), and D the sea storm duration (in hours). Under the assumption that $\mathbf{X}$ can be modeled by an EV or an Archimedean copula, the main idea is to provide a semi-parametric approximation of such a copula, avoiding the fit of any specific parametric model. The (highly non-linear) structure function Ψ in Eq. (1) is calibrated for the buoy of Alghero (Sardinia, Italy), as illustrated in [11], Section 3, where all the parameters (water density, ρ_W, cube density, ρ_S, number of units displaced, N_d, gravitational acceleration, g) are specified:

$$W = \rho_S \cdot \left[H \left(\frac{2\pi H}{g\,[4.597 \cdot H^{0.328}]^2} \right)^{0.1} \right]^3 \Big/ \left[\left(\frac{\rho_S}{\rho_W} - 1 \right) \cdot \left(1 + \frac{6.7 \cdot N_d^{0.4}}{(3600 D / [4.597 \cdot H^{0.328}])^{0.3}} \right) \right]^3 . \quad (2)$$

As marginal distributions for H and D, we adopt suitable Generalized Weibull laws. Via the Monte Carlo strategy sketched in [12], suitable pairs (H, D)'s can be simulated from the approximating copula, and corresponding sample values of W can be calculated, yielding an estimate of the empirical distribution function of W. Then suitable design values w_T's for W, corresponding to specific RP T's, can be computed via the standard formula

$$w_T = F_W^{-1}(1 - \mu/T) \quad (3)$$

where F_W is the law of W, and $\mu > 0$ is the mean inter-arrival time between successive sea storms.

2 Semi-parametric approximations

In order to provide a suitable approximating copula, two characterizing functions known as Pickands dependence function and Kendall distribution function are used in the EV and Archimedean cases, respectively. We recall that a bivariate copula $\mathbf{C}$ is simply (the restriction of) a joint distribution over $\mathbf{S} = [0,1] \times [0,1]$, whose univariate marginals are Uniform. A 2-copula $\mathbf{C}$ is bivariate EV ([5]) if there exists a Pickands dependence function $\mathbf{A}$ such that $\mathbf{C}(u,v) = \exp(\ln(uv)\,\mathbf{A}(\ln v / \ln(uv)))$, for all $(u,v) \in \mathbf{S}$, being $\mathbf{A}\colon [0,1] \to [1/2, 1]$ a convex function satisfying the constraint $\max\{t, 1-t\} \le \mathbf{A}(t) \le 1$, for all $t \in [0,1]$. The uni-dimensional function $\mathbf{A}$ uniquely identifies $\mathbf{C}$. Also, the Kendall distribution function $\mathbf{K_C}\colon [0,1] \to [0,1]$ associated with $\mathbf{C}$ is defined as $\mathbf{K_C}(t) = \mathbf{P}(\mathbf{C}(U,V) \le t)$, where $t \in [0,1]$ is a probability level, and U, V are Uniform random variables on $[0,1]$ with 2-copula $\mathbf{C}$ ([2]). Then, the generator γ of a suitable Archimedean copula $\mathbf{C}(u,v) = \gamma^{[-1]}(\gamma(u) + \gamma(v))$ can be expressed in terms of the Kendall's function $\mathbf{K}$ associated with $\mathbf{C}$ via the formula $\gamma(t) = \exp\left(\int_{t_0}^{t} 1/(x - \mathbf{K}(x))\,dx\right)$, with $t, t_0 \in (0,1)$ and t_0 arbitrary. The equivalent differential form is provided by [8], Theorem 4.3.4. Then, $\mathbf{C}$ can be taken as the representative element of the class of copulas sharing the same function $\mathbf{K}$.
Now, the approximating technique consists in two steps: first, fit a suitable auxiliary function (the Pickands and the Kendall functions, respectively) to the available data, and, secondly, construct an appropriate copula generator exploiting such an auxiliary function. It is worth stressing that a piecewise linear interpolation scheme (both for the EV and the Archimedean case) is adopted, since it represents

a natural and basic semi-parametric approximation to the generators of interest. In addition, the convergence of such linear approximations is guaranteed under minimal conditions, which, in turn, yields the convergence of the corresponding copulas ([10, 9]).

2.1 The EV case

Let $\mathcal{X}_n$ denote a set of abscissas $x_0 = 0 < x_1 < \cdots < x_{n-1} < x_n = 1$ in $\mathbf{I} = [0,1]$, with $n \in \mathbf{N}$. Moreover, $\mathbf{X}$ will denote a sample of size $N > 0$ of i.i.d. bivariate sea storms (H,D)'s.
A possible procedure for supplying approximate structural estimates is as follows. (i) Provide an estimate $\mathbf{A}_n$ of the Pickands dependence function $\mathbf{A}$ associated with $\mathbf{X}$ at the points of $\mathcal{X}_n$ via some common rank-based estimator (available in the R-package "Copula" [7]). (ii) Generate two independent variates u and w Uniform over $(0,1)$. (iii) Numerically compute $v = c_u^{-1}(w)$ where the function $c_u(v) = \partial \mathbf{C}_n(u,v)/\partial u$ can be calculated by, first, computing the approximating copula $\mathbf{C}_n$ associated with $\mathbf{A}_n$, and then applying suitable numerical formulas for estimating the partial derivative. Finally, the pair $\left(H = F_H^{-1}(u), D = F_D^{-1}(v)\right)$ can be obtained by inversion of marginal laws and Eq. (2) used to fix the corresponding cube weight W.
Once a suitable sample of simulated cube weights $\mathbf{W}$'s is made available, Eq. (3) can be used to provide approximate design values w_T's, for prescribed RP's T's, by inverting the empirical distribution function of the $\mathbf{W}$'s. It is possible to show that $\mathbf{A}_n$ is a consistent and asymptotic Normal estimator of $\mathbf{A}$ ([3]). Thus, the corresponding approximating copula $\mathbf{C}_n$ provides a consistent estimator of the true copula $\mathbf{C}$.

2.2 The Archimedean case

Let $\mathcal{K}_n = \{k_0 = 0 \le k_1 \le \cdots \le k_{n-1} \le k_n = 1\}$ be a set of estimates of the Kendall distribution function computed at the points of $\mathcal{X}_n$. Given the set of points (x_i, k_i), with $i = 0, \ldots, n$, the procedure is as follows. (i) Generate two independent variates s and t Uniform over $(0,1)$. (ii) Numerically calculate $w = \gamma_n(\mathbf{K}_n^{-1}(t))$, where an approximation of the Archimedean generator γ_n is computed by using suitable numerical integration procedures. (iii) Numerically compute $u = \gamma_n^{-1}(s\,w)$ and $v = \gamma_n^{-1}((1-s)\,w)$.
Finally, use the marginal laws to calculate the pair $\left(H = F_H^{-1}(u), D = F_D^{-1}(v)\right)$, and the expression for W given by Eq. (2) to fix the corresponding cube weight W. Approximate design values w_T's, for prescribed RP's T's, can be obtained by inverting the empirical distribution function of the sample of simulated cube weights $\mathbf{W}$'s (Eq. (3)). It is possible to show that $\mathbf{K}_n$ is a consistent and asymptotic Normal estimator of $\mathbf{K}$ ([6]). In turn, the copula $\mathbf{C}_n$ associated with $\mathbf{K}_n$ is a consistent estimator of $\mathbf{C}$ (see also [10]).

3 Monte Carlo tests

In order to test the performance of the two approaches outlined above, the Gumbel family of copulas is chosen, since it is both EV and Archimedean. The evaluation of the performances is based on the response of the structure considered, and on the comparison of the estimates with the "true" Gumbel values that arise when the model is known. In particular, the values corresponding to different design quantiles of W are considered, corresponding to standard RP's of 10, 20, 50, 100 years. A sample of sea storms of size N is generated, and used to provide approximate design values of the cube weights. The latter ones are then compared with the "true" Gumbel ones, with Kendall's rank correlation coefficient τ. By

varying N and τ, it is possible to draw fairly exhaustive statistical scenarios concerning the performance of the EV and Archimedean strategies introduced. Overall, the results (not reported here) indicate that the proposed approach may provide valuable estimates of the sought quantiles, without making recourse to any specific parametric model.

4 Conclusions

The procedures outlined in this work are essentially based on the fact that both EV and Archimedean copulas may be suitable to describe the joint statistical behavior of the variables involved in many environmental applications. One main advantage is that they can be generated by using auxiliary one-dimensional functions (the Pickands dependence function and the Kendall distribution function), and estimated using the available data. The techniques outlined in this work may provide valuable approximations to some (large) Return Period design values, commonly used in risk assessment. Monte Carlo tests and a robustness-sensitivity study (not discussed here) are carried out in order to investigate the performance of the approximating algorithms. As a result, the procedures presented in this work may provide valuable indications for a preliminary assessment of the structural risk and the choice of a suitable parametric model.

References

[1] Durante, F., Okhrin, O. (2015). Estimation procedures for exchangeable Marshall copulas with hydrological application. *Stochastic Environmental Research and Risk Assessment* **29**, 205–226.

[2] Genest, C., Rivest, LP. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* **88**, 1034–104.3

[3] Genest, C., Segers, J. (2009). Rank-based inference for bivariate Extreme Value copulas. *Annals of Statistics* **37**, 2990–3022.

[4] Grimaldi, S., Koutsoyiannis, D., Piccolo, D., Schumann, A. (2009). Recent developments of statistical tools for hydrological application. *Physics and Chemistry of the Earth* **34**, 595–595.

[5] Gudendorf, G., Segers, J. (2010). Extreme-value copulas. In: Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds). *Copula Theory and its Applications, Lecture Notes in Statistics - Proceedings* **198**, 127–145. Springer. Berlin Heidelberg.

[6] Hofert, M., Mächler, M., McNeil, AJ. (2012). Likelihood inference for Archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis* **10** 133–150.

[7] Hofert, M., Kojadinovic, I., Maechler, M., Yan, J. (2013). Copula: Multivariate Dependence with Copulas. R package version 0.999-7 edn.

[8] Nelsen, R. (2006). *An introduction to copulas*. 2nd edn. Springer-Verlag, New York.

[9] Pappadà, R. (2009). *Copule bivariate dei Valori Estremi*. Tesi di Laurea Specialistica (in Italian), Dipartimento di Matematica, Università del Salento, Lecce (Italy).

[10] Salvadori, G., Durante, F., Perrone, E. (2013). Semi-parametric approximation of the Kendall's distribution and multivariate Return Periods. *Journal de la Société Française de Statistique* **154**, 151–173.

[11] Salvadori, G., Durante, F., Tomasicchio, G.R., D'Alessandro, F. (2015). Practical guidelines for the multivariate assessment of the structural risk in coastal and off-shore engineering. *Coastal Engineering* **95**, 77–83.

[12] Straub, D. (2014). Engineering risk assessment. In: Klüppelberg, C., Straub, D., Welpe, IM. (eds) *Risk - A Multidisciplinary Introduction* 333–362. Springer International Publishing Switzerland.

Impact of climatic factors on acute bloody diarrhea, dengue and influenza-like illness incidences in the Philippines

A. Rarugal[1,*], R. M. Roxas-Villanueva[2] and G. Tapang[1]

[1] *National Institute of Physics, University of the Philippines Diliman, Quezon City 1101, Philippines; ararugal@nip.upd.edu.ph, gtapang@nip.upd.edu.ph*
[2] *Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, Laguna 4030, Philippines; rrvillanueva3@up.edu.ph*
[*] *Corresponding author*

Abstract. *The effect of climate variability on the weekly incidence of acute bloody diarrhea, dengue and influenza-like illness in the 17 regions of the Philippines is examined using correlation, mutual information and transfer entropy. Results show that the correlations between climate variables and disease incidences differ from one region to another. Interestingly, the diseases are directly correlated to each other for each region. This is explained by their common driving climate factors which are shown by large transfer entropy values. This work is important in further understanding the role of climate variability to the temporal dynamics of disease incidences.*

Keywords. *Time series analysis; Statistics; Information theory*

1 Climate Change and Human Health

Climate change affects human health. Several endemic human diseases are linked to climate variability from cardiovascular and respiratory disease fluctuations influenced by heatwaves, transmission of communicable diseases by adaptation of vectors, to malnutrition caused by crop shortage [7, 8]. Philippines is the third most disaster-prone nation in the world. It is susceptible to many environmental upshots due to active volcanoes, rich biodiversity and regular typhoon occurrences [9]. It is therefore important to examine the effects of climate variability on health. Evidence and expectancy of the harmful effects will improve preemptive policies and adaptive solutions [3].

This work investigates the impact of climate variables on acute bloody diarrhea, dengue and influenza-like illness in the 17 regions of the Philippines using time series analysis. Statistical and information theoretic measures known as correlation, mutual information and transfer entropy are used to determine important relationships between the climate variables and disease incidences.

2 Methodology

The National Epidemiology Center (NEC) of the Philippines [5] provided the number of reported cases of acute bloody diarrhea, dengue and influenza-like illness from January 1, 2012 to May 4, 2013. The climate variables include average, maximum and minimum air temperature, diurnal temperature range (°C), relative humidity (%), dew or frost point temperature (°C), rainfall (mm/day) and wind speed (m/s). These daily data, obtained from National Aeronautics and Space Administration Prediction of Worldwide Energy Resource (NASA POWER) [4], are converted to mean weekly time series to be consistent with the frequency of the disease data.

Three techniques in time series analysis are applied to evaluate the impact of climate variables on the disease incidences. For two time series X and Y with measured observations $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$, respectively, the correlation coefficient r or Pearson's r determines the extent of how X and Y follow each other's path through time [1]. It is a measure of linear association but does not immediately imply causality [10]. Moreover, mutual information measures the similarity between the pair of time series by measuring the amount of information common to them. Unlike Pearson's r, mutual information can detect nonlinear dependencies. It is mathematically expressed as

$$I(X,Y) = \sum_i p_{X,Y}(x_i, y_i) \log \left(\frac{p_{X,Y}(x_i, y_i)}{p_x(x_i) p_y(y_i)} \right) \tag{1}$$

where $p_{X,Y}(x,y)$ is the joint probability distribution of X and Y and $p_x(x)$ and $p_y(y)$ are the corresponding marginal probabilities [1]. Finally, transfer entropy determines the magnitude and directionality of information exchange between Y and X. Such tool is non-symmetric under information exchange which makes it capable to quantify the influence of Y on the evolution of X [2]. It is given by

$$T_{Y \to X} = \sum_k p\left(x_{k+1}, x_k, y_k\right) \log_2 \frac{p\left(x_{k+1} \mid x_k, y_k\right)}{p\left(x_{k+1} \mid x_k\right)} \tag{2}$$

where $p(x|y)$ are conditional probabilities. It can determine whether two systems are similar because another system influences the first two. [1].

3 Results and Discussion

Figure 1 presents the regional map of the Philippines and the spatial distribution of the mean weekly number of the diseases in the country. On the average, there were 46 reported acute bloody diarrhea cases, 83 dengue cases and 7 influenza-like illness cases nationwide every week from January 1, 2012 to May 4, 2013. Region III (Central Luzon) has the lowest mean number of acute bloody diarrhea cases which suggests that it is the least vulnerable region to the said disease. On the other hand, Region II (Cagayan Valley) has the highest mean number of cases. Region X (Northern Mindanao) and Region IX (Zamboanga Peninsula) are the least and most vulnerable regions to dengue, respectively. Furthermore, Region III is the least vulnerable to influenza-like illness while CAR (Cordillera Administrative Region) is the most vulnerable.

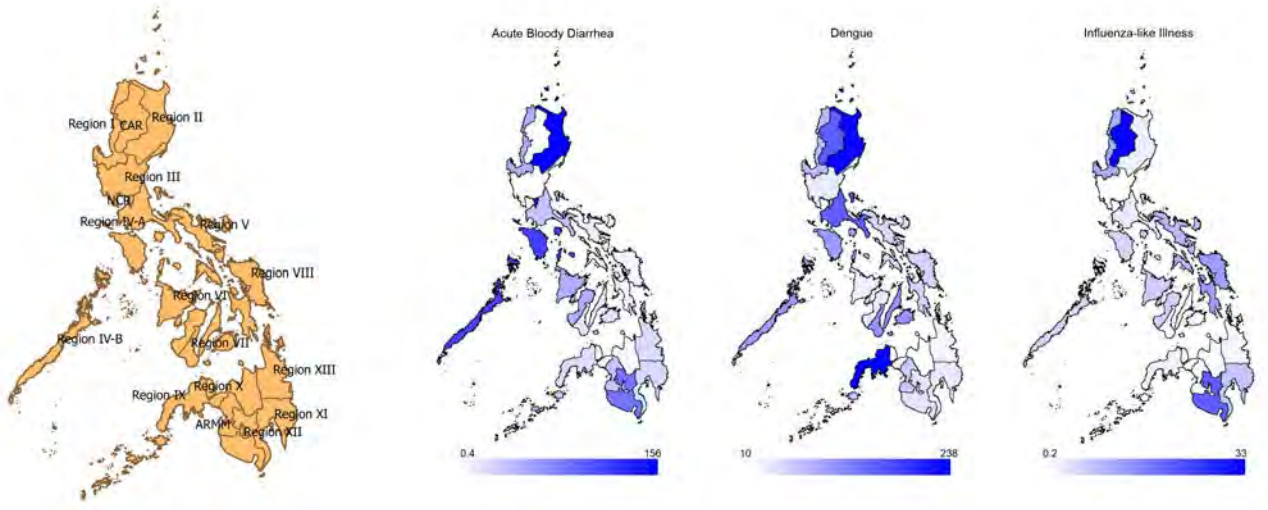


Figure 1: Regional map (orange) of the Philippines and mean weekly number of acute bloody diarrhea, dengue and influenza-like illness reported cases (blue) in the country.

Figure 2 summarizes the temporal associations between the diseases and climate variables in Region I by presenting the correlation, mutual information and transfer entropy values between the datasets. Correlation is significant when $|r| > 0.2352$ by two-tailed student's t-test at 0.05 level of significance [6, 10]. From Figure 2a, dengue is inversely correlated to maximum temperature and changes in temperature and is directly correlated to humidity, dew or frost point temperature and rainfall fluctuations in Region I. It can also be observed from Figure 2b that acute bloody diarrhea is not mutually dependent to any climate variable. On the other hand, dengue, influenza-like illness and all the climate variables are mutually dependent despite the fact that majority of these relationships are not detected by linear correlation.

Information transfer patterns by transfer entropy imply that acute bloody diarrhea, although not correlated with any climate variable, together with dengue are driven by changes in temperature as shown by their large transfer entropy values. It can also be observed that there are significant linear correlations between the three diseases. Detected correlations can therefore be explained by their common driving elements.

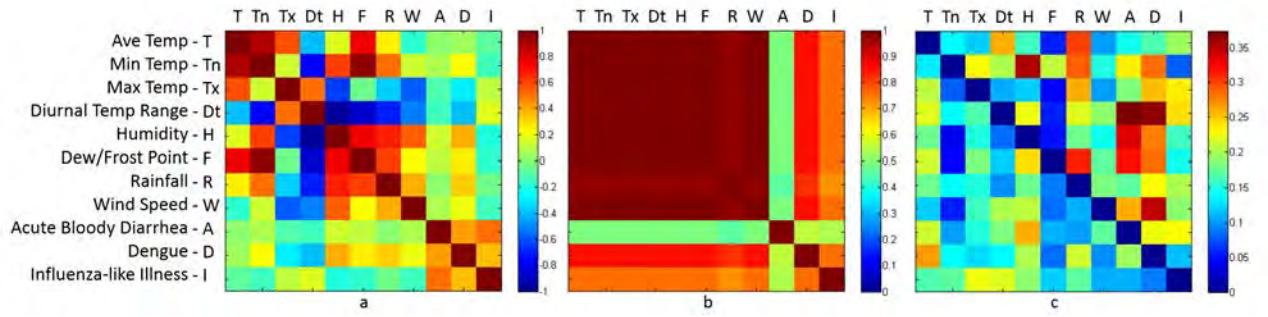


Figure 2: (a) Correlation, (b) mutual information and (c) transfer entropy between the climate variable and diseases in Region 1 (Ilocos Region).

The corresponding correlation, mutual information and transfer entropy plots of the remaining 16 regions were also generated. In general, the three diseases are directly correlated to each other for each region. The linear and nonlinear associations can be explained by the common driving climate variables. Moreover, the relationships between the diseases and climate variables differ from one region to another. Such dissimilarities can be accounted for the differences in geography, population and other features of the regions of the Philippines.

4 Conclusion

The extent of the effects of climate variables on the incidence of acute bloody diarrhea, dengue and influenza-like illness vary from one region to another as manifested by the calculated correlation, mutual information and transfer entropy values. However, linear correlations between the diseases are consistently observed for each region. The dissimilarities can be accounted for the disparate geography, population and other attributes of each region. The linear and nonlinear associations, on the other hand, can be explained by the common driving climate variables. Improving the length and the resolution of the datasets could reveal more information about the relationship between the diseases and climate variables. This work is important in further understanding the role of climate variability to the temporal dynamics of disease incidences. It can be used to mitigate disease outbreaks and to employ practical solutions to climate-related health problems.

Acknowledgments. The researchers would like to thank NASA Langley Research Center POWER Project funded through the NASA Earth Science Directorate Applied Science Program for the climate data and NEC for the disease data.

References

[1] Albano, A., et. al. (2008). Time series analysis, or the quest for quantitative measures of time dependent behavior. *Philippine Science Letters* **1**, 18–31.

[2] Kaiser, A. and Schreiber, T. (2002). Information transfer in continuous processes. *Physica D* **166**, 43–62.

[3] McMichael, A. J., et. al. (2006). Climate change and human health: present and future risks. *The Lancet* **367**, 859–869.

[4] NASA POWER Climatology Resource for Agroclimatology. Available at http://www.power.larc.nasa.gov/.

[5] National Epidemiology Center Weekly Disease Surveillance Report. Available at http://www.nec.doh.gov.ph/.

[6] NIST/SEMATECH e-Handbook of Statistical Methods. Available at www.itl.nist.gov/div898/handbook/.

[7] Patz, J. A., et. al. (2005). Impact of regional climate change on human health. *Nature* **438**, 310–317.

[8] Smith, K. R., et. al. (2014). *Human health: impacts, adaptation, and co-benefits. In Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate*. Cambridge University Press. Cambridge, United Kingdom and New York, New York.

[9] United Nations Philippines. Country Profile. Available at http://www.un.org.ph/country-profile.

[10] Witte, R. S. and Witte, J. S. (2001). *Statistics*. Harcourt, Inc. Orlando, Florida.

Official statistics for decision making: an environmental accounting case study related to biodiversity

E. Recchini[1]

[1] *Istat (Italian National Institute of Statistics); recchini@istat.it*

Abstract. *Decision-makers' increasing awareness of biological resources' importance for humanity's economic and social development has driven the inclusion of biodiversity among the main topics dealt with by official statistics, thus enabling an extension of traditional analyses based on social and economic data to cover crucial environmental and sustainability aspects. Within official statistics, environmental-economic accounts can play a specific role in supporting initiatives stemming from the UN Convention on Biological Diversity (CBD). Together with core national accounts, environmental accounting is involved even directly in the implementation of the Aichi Biodiversity Targets (ABTs) agreed upon in the context of the Strategic Plan for Biodiversity 2011-2020: according to ABT2, biodiversity values are to be incorporated into reporting systems and into national accounting as appropriate. In particular, data for monitoring the mobilization of financial resources for the implementation of this Strategic Plan as well as for assessing resource needs are of interest according to ABT20. For these purposes, data derived from official statistics' environmental accounting on actual expenditure for biodiversity protection have special merits, due to their high quality and to the fact that they are linked to core national accounts data and hence particularly suitable for use in modelling. On the other hand, while policies can easily be developed by committing financial flows to given purposes, monitoring the same policies based on information on actual utilization of money may not be that easy from politicians' viewpoint. The use of data on funding vs data on actual expenditure may be an issue.*

Keywords. *Environmental-economic accounts; environmental protection expenditure; protection of biodiversity.*

1 Introduction

Statistical information is crucial for decision making: it enables to identify key areas where actions are required and, if correctly interpreted, allows decision-makers to respond to the real needs of a community. Statistics can also help the general public to monitor and evaluate the performance of politicians and decision-makers.

On the impulse of a wide-spread concern about human impacts on ecosystems and biological diversity, the demand for statistics on phenomena related to biodiversity has been increasing during the last decades.

Biological diversity has emerged as a fundamental part of the move towards sustainable

development, and its conservation and the sustainable use of its components are on the diplomatic agenda. The United Nations adopted in 1992 the Convention on Biological Diversity (CBD), the first global, comprehensive agreement to address all aspects of biological diversity. Inspired by the world community's growing commitment to sustainable development, CBD recognizes, for the first time, that the conservation of biological diversity is "a common concern of humankind" and an integral part of the development process. Given a general lack of information and knowledge regarding biological diversity, CBD highlights the urgent need to develop scientific, technical and institutional capacities to provide the basic understanding upon which to plan and implement appropriate measures. For the implementation of CBD the Strategic Plan for Biodiversity 2011-2020 (SPB) is currently in place, providing goals and specific targets aimed to better preserve and protect natural resources with sustainable management. One important process in this context is known as "Biodiversity Resource Mobilization" (BRM). Of course, there is a need of statistical information suitable for the purposes of CBD, including BRM. The so-called Aichi Biodiversity Targets (ABTs), agreed upon in this context, include this aspect as well: according to ABT2, biodiversity values are to be incorporated into reporting systems and into national accounting as appropriate; in particular, data are needed for monitoring the mobilization of financial resources for the implementation of the strategic plan as well as for assessing resource needs (ABT20[1]).

Biodiversity is being taken into account more and more not only from an ecological viewpoint but also as far as related economic and social aspects are concerned. There is a growing consensus that biodiversity is fundamental to economics, as witnessed e.g. by the global initiative TEEB[2]. A specific kind of information which is relevant to support action for the conservation of biodiversity is then one that links ecological and economic aspects. Suitable data for such linkages is provided by environmental-economic accounts.

2 Environmental-economic accounts: an international statistical standard with a legal base in the EU

Within official statistics the interaction between economy and environment is described by means of accounts that are satellites to the national accounts as well as other statistics that consider both environmental and economic aspects at the same time. Environmental-economic satellite accounts link the environmental and economic dimensions based on a system approach. This is done according to SEEA[3], an overarching international framework based on the same basic principles, definitions and classifications of the core system of national accounts, thus allowing proper linkages with economic accounting data[4]. SEEA is a framework to organize data for the derivation of coherent indicators and descriptive statistics to monitor the contribution of the environment to the economy and the impact of the economy on the environment, as well as the state of the environment. It is developed in a way that the different domains of the environmental debate can be suitably covered by statistics produced according to its guidelines. Biodiversity is one of such domains; one specific SEEA module provides data on expenditures for the protection of biodiversity, which is of particular importance in relation to ABT20 mentioned above.

The UN Statistical Commission (UNSC) endorsed SEEA following a request from Agenda 21[5]. In

[1] ABT20 reads: "By 2020, at the latest, the mobilization of financial resources for effectively implementing the Strategic Plan for Biodiversity 2011-2020 from all sources, and in accordance with the consolidated and agreed process in the Strategy for Resource Mobilization, should increase substantially from the current levels. This target will be subject to changes contingent to resource needs assessments to be developed and reported by Parties.".
[2] http://www.teebweb.org/.
[3] System of System of Environmental-Economic Accounting 2012.
[4] SNA 2008 (http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf).
[5] UN Agenda 21, the action plan adopted at the Earth Summit held in Rio de Janeiro in 1992 for implementation worldwide, is at the origin of SEEA. It was Agenda 21 that called for the development of integrated environmental-economic accounting from different perspectives, including official statistics as well as corporate reporting.

particular, the SEEA Central Framework (SEEA-CF) - one component of SEEA - has been adopted by UNSC as an international statistical standard, similarly to SNA, the above mentioned core system of national accounts. Thanks to the consistency of SEEA with SNA, the trade-offs of policy-makers' economic decisions affecting natural resources and associated services can be made explicit.

SEEA Experimental Ecosystem Accounting (SEEA-EEA) - another component of SEEA[6] - deals with biodiversity aspects more in depth as compared to SEEA-CF. The former is not an international standard, but complements the latter by providing methodological guidelines for accounts focused on ecosystems. It is two modules of SEEA-CF, however, that provide proper accounting tools to describe the relevant expenditures: EPEA[7] and ReMEA[8]. These two modules are being systematically produced in most EU member countries.

In fact, a legal base has been established in EU for mandatory production of national environmental-economic accounts in line with SEEA: Regulation of the European Parliament and of the Council on European environmental economic accounts (No 691/2011), amended by Regulation No 538/2014[9]. This legal base provides methodology, common standards, definitions, classifications and accounting rules for compiling accounts that are given highest priority in EU according to the European Strategy for Environmental Accounts (ESEA)[10]. The second Regulation mentioned above provides the legal base for the module "Environmental protection expenditure accounts", which is particularly relevant for the calculation of economic resources devoted by resident units to environmental protection, including for conservation of biodiversity.

3 Monitoring conservation of biodiversity: data on funding vs data on actual expenditure

With connection to the EU Biodiversity Strategy to 2020[11], EU member countries have developed a system of indicators – namely the Streamlining European Biodiversity Indicators (SEBI) - which are supposed to be used for monitoring the implementation of the same strategy and, in a global perspective, also the attainment of the corresponding ABTs[12].

In the above context, one indicator is intended to assess how much public funds are being committed to conservation of biodiversity: SEBI 025 - "Financing biodiversity management". One main limit of this indicator has emerged in practice: it only contains information from EU funding; furthermore only funding of projects using the LIFE financial instrument for the environment is considered, while European funding benefiting biodiversity may also be included, though not explicitly, in budget lines within other policy areas, e.g. agriculture, rural development and research.

Beyond shortcomings which any given indicator may show at a given stage of its development, the case of SEBI 025 suggests a reflection on which kind of data could serve as an optimal indicator for monitoring expenditure, in particular for conservation of biodiversity as in the case at study. In general terms the issue is whether, for monitoring decision making processes, data on funding would be better suited than data on actual expenditure or vice-versa.

[6] One more component of SEEA is Applications and Extensions (of the SEEA). So-called subsystems of the SEEA framework have also been developed in order to elaborate on specific resources or sectors, e.g. Energy, Water.
[7] Environmental Protection Expenditure Account.
[8] Resource Management Expenditure Account. In SEEA-CF this module is envisaged, though not actually developed in operational terms.
[9] This regulation adds three new modules to those initially introduced by the first Regulation in 2011.
[10] The legal base is supposed to be further extended to cover more modules, also in accordance with ESEA.
[11] http://ec.europa.eu/environment/nature/biodiversity/comm2006/2020.htm.
[12] In fact, the indicators at issue are aimed to link the global framework set by the CBD with regional and national indicator initiatives.

Italy is a country extremely rich in biodiversity - due to its territory with remarkable differences in climate, topography and geology - and is strongly committed to the implementation of its strategies related to biodiversity: National Biodiversity Strategy and National Strategy for Resource Mobilization, both linked to CBD and related SPB and ABT20 mentioned above. Estimates on actual national expenditure for conservation of biodiversity are regularly produced in Italy by Istat according to EPEA and ReMEA mentioned above. Two specific environmental domains covered in these accounts are relevant in relation to ABT20: "Protection of biodiversity and landscapes" and "Management of wild flora and fauna"[13]. Like all figures delivered by Istat, expenditure estimates referred to these domains are produced in compliance with the European Statistics Code of Practice[14].

In principle, these official statistics can be used for assessing financial resource needs as well as for calculating the resources made available, in order to support studies useful for the purposes of CBD, if not to contribute to the assessments of economic efforts carried out for the conservation of biodiversity implicitly required by ABT20. However, also due to the fact that actual national expenditure is not calculated all over the world on a regular basis and consistently with EPEA, this kind of data does not seem to enter international negotiations for conservation of biodiversity. On the other hand it may also be relevant that, while policies can easily be developed by defining the financial flows to be committed to given purposes, from politicians' viewpoint it may be not equally straightforward to monitor the same policies by using information on actual utilization of money.

4 Concluding remarks

As in other international processes, there may be good reasons in the "Biodiversity Resource Mobilization" process for understanding resource mobilization just as funding. At least, policies can easily and significantly be developed by committing financial flows to given purposes, while it is less easy to monitor their implementation and effectiveness in terms of actual expenditure. However, actual utilization of the financial resources committed to CBS's purposes does matter as well, because eventually efforts and activities actually carried out to protect biodiversity is what really matters.

The usefulness of official statistics derived from SEEA and SNA seem to be out of discussion, nevertheless, at least because national accounting is mentioned in ABT2. Such statistics, furthermore, would be useful for assessing resource needs as politicians might want to take an informed decision to change ABT20 as envised in the target itself.

References

[1] SEEA (2014), System of Environmental-Economic Accounting 2012 – Central Framework. United Nations, New York (http://unstats.un.org/unsd/envaccounting/seeaRev/SEEA_CF_Final_en.pdf).

[2] European Commission, International Monetary Fund, OECD, United Nations, World Bank (2009). System of National Accounts 2008, New York.

[3] United Nations (1992), Convention on Biological Diversity (http://www.cbd.int/).

[4] The Economics of Ecosystems and Biodiversity (http://www.teebweb.org/).

[13] "Protection of biodiversity and landscapes" is item 6 of the European standard statistical Classification of Environmental Protection Activities and Expenditure (CEPA), used in EPEA and ReMEA and adopted in the above mentioned Regulation No 538/2014. "Management of wild flora and fauna" is item 12 of the Classification of Resource Management Activities (CReMA), used in ReMEA and adopted in the same Regulation for compiling statistics on the Environmental Goods and Services Sector.

[14] Mandatory quality assurance procedures are regularly carried out.

Environmental sustainable management of urban networks with the use of ICT: URBANETS project. The case of Gallipoli

E. Venezia[1]

[1]*University of Bari Aldo Moro, Department of Economics and Mathematical Methods, Largo Abbazia Santa Scolastica, 70124 Bari, Italy, elisabetta.venezia@uniba.it*

Abstract. *This paper is part of a wider work – developed in the context of URBANETS project - Sustainable Management of Urban Networks with the Use of ICT – on traffic and environmental requirements for Brindisi and Gallipoli. The paper considers different aspects, having as an objective the satisfaction of transport and environmental knowledge needs with particular regard to Gallipoli. The aim is to supply policy indications in the light of population, economic operators and stakeholders exigencies. The paper contains transport and environmental analyses referred to Gallipoli area and results, stemming from a consultation process of transport operators, citizens, tourists, public employees and stakeholders, are presented. Through the statistical analysis of data (which describes the nature of data, explore the relation of the data to the underlying population, create a model and prove its validity), here shortly presented, several insights are provided on needs of specific user categories, as gender issues and social equity are key issues in urban policies. In addition, the potential willingness to pay of users to obtain a general improvement in bus service quality and in environmental conditions is investigated through discrete choice modelling. The idea behind this study is to overcome the crucial impediment in understanding urban travel patterns and the key forces behind user attitudes which normally characterise city surveys. Therefore, attitudinal and behavioural variables are considered to evaluate the propensity of using buses and changing habits for modal choices, more environmental sustainable, through a random utility model. Finally, indications on perspectives and final conclusions are supplied. This paper can represent a useful tool for those who operate in the transport and environmental sectors and for policy-makers as well.*

Keywords. *Sustainable mobility; Environmental impact; Random utility model; Willingness to pay; Policy indications.*

1 Results from consultation process of stakeholders and focus groups operating in the transport sector

For the city of Gallipoli some firms have been selected in the data base of Italian Chamber of Commerce. They are part of the transport sector with the ATECO codes H49, H50, H51 e H53. These firms have been contacted to understand the main critical points for the territory under investigation, particularly with reference to the transport sector. The idea is to have a framework of the structure of freight transport supply – also with regard to the location – in the light of territorial characteristics. Firms selection for running focus groups, realised in October and November 2013, were 192. Questionnaire was intentionally brief so that characteristic elements of transport supply could immediately emerge. This aspect contributed to collect useful information for an optimal choice related to transport market organization and logistic services. As for stakeholders, they have been asked

specific questions related to transport and environmental aspects.

1.1 Consultation results

On the basis of contributions obtained, it is possible to say that the current transport system is not considered adequate because the seasonal phenomenon is still heavily hitting transport and environmental situation. Furthermore, there is a problem linked to the funnel shape of the access to the historical centre which creates enormous congestion problems, especially in the summer when there is also the need to have access to the sea. Therefore, a major problem is linked to tourist traffic which have seasonal effects. There is a mismatch demand/supply with regard to private motorized modes and the available supply. Furthermore, road network is not considered adequate by interviewees as for private means and buses, and there are also problems of road burden capacity and of tourist reception. Public services are supplied in accordance with the service contract, nevertheless there are opportunities to cover a higher geographical area and to extend timing for supplying the service. In fact the demand is not satisfied with regard to these two factors. Suggested improvements to overcome criticalities are of infrastructural type and linked to a demand regulation with the closer to traffic of congested areas. It is required an alternative service supply more adequate to users' needs. Effects of the suggested interventions on the territory and on the environment are indicated as a better accessibility and an improved environment as a consequence of different modal choices in favour of more environmental sustainable transport means. To finance those interventions it is asked the participation of the Municipality of Gallipoli and of private investors, where this is possible. Among the priority transport infrastructures, interviewees indicate the construction of a new coastroad, parallel to the existing one and the installation of monitoring control units for the air quality to check transport impact. So doing it could possible to intervene with appropriate transport demand management tools and to regulate the access to congested areas. Suggestions also indicate to realize infrastructures to favour bike mobility.
To conclude, it is possible to say that from the realized analysis and from the desk analysis of available data and those stemming from consultations, a clearer framework for Gallipoli has been obtained with regard to needs and possible solutions. Therefore it is clear that this area has a high level of transport demand in comparison with the regional framework, which is satisfied almost exclusively with only one mode: road transport. As a consequence operators and stakeholders, on one side, ask for reducing congestion in these areas with infrastructural investments, economic interventions and laws in order to manage transport demand. On the other side, it is useful to intervene with tools to modify modal choices and address them to environmentally-friendly choices. Interviewees also say that in order to stimulate local development is indispensable to realise the suggested interventions and supply adequate services in line with the expressed needs. So doing opportunity of growth can occur and it can attract further tourism which is a competitive leverage for the territory. Besides, also transport needs can be satisfied and the city can offer a better accessibility through an efficient transport system.

2 Survey results and empirical findings on the WTP

In the months of November and December 2013 in the city of Gallipoli interviews have been done to 383 persons next to crucial transport points of the city, in some hotels and in front-offices of the Municipality of Gallipoli. Moreover, also employees of the Municipality of Gallipoli took part to this consultation. The questionnaire was structured by considering a personal profile (like age, gender and occupational status) modal transport choices and motivations, evaluation of transport modes, and perspective preferences and willingness to pay. In this sample, with regard to the gender aspect, there are few additional female in comparisons with males. It comes out that females are 53,3% while males are 46,7%. The age composition of this sample is structured with more than 35% respondents in the range 30-50, followed by the range 19-29 years old, by the range 51-65 years old, then by respondents under 19 and, finally, by those over 65 years old with 8%. With regard to those who have the availability of their own cars, results indicate that 45,69% of respondents have always a private mean, while those who can use a private mode only sometime are 41,78% of the total sample. Finally, with a low percentage equal to 12,53%, there are those who do not have a private mean at all. This is a clear indication of how private cars are normally used in a family context. With regard to results related to transport modes used by interviewees, it emerges a high percentage of those using private cars, equal to 38,62%. To this figure must be added also motorcycles and bicycles which recorded respectively 6,65%

and 4,6%. Interesting for sustainable considerations is the percentage of those who move by foot equal to 12,53%. This is justified by the fact that Gallipoli is a small city. Finally, there are also those who use taxi . This component is referred to tourists and business movements. As for the frequency in the use of buses, it is possible to highlight that 30,4% of interviewees has declared to use only rarely buses. This percentage, considered in combination with that referred to respondents that never use buses (20,1%), gives a clear signal of modal choices and the lack of consideration of public transport for internal movements. There is also 22,4% of respondents who always use buses, followed by those who get buses 1-2 times a week (21%) and 3-4 times a week (6,1%).

To give a further interpretation of data on individual choice related to bus service supplied in Gallipoli, a random utility model framework has been used. As indicated by Green (1997) [1], suppose that ym and yp represent the individual's utility of two choices, denoted Ua e Ub . The observed choice between the two reveals which one provides the greater utility. Therefore, the observed indicator equals 1 if Ua > Ub and 0 if Ua ≤ Ub . A common formulation of the linear random utility model is:

Ua = ß'a x + ε a and Ub = ß'b x + ε b.

Then if we denote by Y=1 the consumer's choice of alternative a, we have:

$$\begin{aligned}\text{Prob}[Y=1|x] &= \text{Prob}[Ua > Ub] \\ &= \text{Prob}[ß'a\ x + \varepsilon a - ß'b\ x - \varepsilon b > 0|x] \\ &= \text{Prob}[(ß\ a - ß\ b)'\ x + \varepsilon a - \varepsilon b > 0|x] \\ &= \text{Prob}[ß'x + \varepsilon > 0|x].\end{aligned}$$

This model is one of the most used for the simulation of transport demand, nevertheless it may present some problems. On this point see Cascetta, E.-Papola, A. (2001) [2], Maddala, G.S. (1999) [3], Green, W.H. (1997) [1]. The individual's utility of two choices – use of buses and use of private means – is estimated by binary logistic regression and logistic regression coefficients are used to estimate odds ratios for each independent variable in the model. The values assumed by the dependent variable, as the probability to use buses, is posed equal to 1. All the values assumed by independent variables have been transformed into dummy variables in order to capture each characteristic of independent variables represented by age, availability of private transport means and so on. Equations have been estimated by using single attribute to avoid evident correlation problems and a consequent self-selectivity involved in the data. Here, the selection is given by the significance of parameters, which has been checked with the Wald statistic at a 5% level. All parameters have been chosen with the Wald forward selection method and values reported in Table 1are all significant in accordance with the Wald test.

Female user profile variable/attribute	Items	Probability to use bus	WTP
Age			
	19-29	8.50	1.55
	30-50	4.10	1.78
	51-65	14.57	
Availability of other transport means			
	Always	3.01	
	Sometimes	0.75	3.12
	Never		1.58
Frequency in the bus use			
	Every day	86.34	1.38
	1-2 times per week	6.37	1.72
	3-4 times per week		3.99
	Rarely	1.51	
Reasons			
	School/working activity	15.62	1.65
	Leisure activity	10.52	1.02
	Shopping	4.93	1.94

Table 1: Women- Probability to use bus and WTP for an improvement in the bus service

Table 1 shows values assumed by coefficients as odds ratios. They indicate the probability of women (the most significant sample section) using a public bus for each characteristic, against the probability of using other means in an urban context, and the willingness to pay something more for a general improvement in the bus service estimated by using the same aforementioned procedure. Probability is outlined as a function of various other profile variables.

Table 1 shows that the probability of females to use buses is particularly important for those belonging to the range 51-65, followed by women with 19-29 years, and those in the range 30-50. For women who have always a car or other transport means, the probability to use a bus is more than twice those who have sometimes a car. The most important reason that can push women to get a bus is for studying or working: it is 3 times more important that those who travel for shopping and one time and a half of those who choose public transport for leisure activity. Finally, Tables 1 contains also information on the willingness of women to pay something more for a general improvement in the bus service which may induce people to change their habits. The procedure followed for parameters selection and estimation methods is exactly the same used to investigate the use of public buses. In particular, for women the best profile is to belong to a range 30-50, to have sometime another transport mean, to take 3-4 times per week a bus, to use the bus service for shopping.

Therefore, results give clear indications on the possibility to satisfy with more sustainable transport collective modes latent systematic transport demand, particularly the one expressed by women, within 19-29 years, who always have a car. At the same time there is the willingness to pay higher public transport tariffs to have a better collective transport service which is, again, largely environmental sustainable.

3 Conclusions

To conclude, results indicate very well intervention needs on the territory of Gallipoli. In fact, although it is appreciable the work already done by the administration in planning and programming terms, it is evident a gap between what population, economic operators and stakeholders ask and what is decided by policy-makers. This is essentially due, on one side, to a delay in the execution of those planes which has the effect of slower operative procedures (which is a common element at national level) and, on the other side, there is not a constant listening of needs so to determine the correct priorities and allocate efficiently the tight financial resources. Also this last element is not only typical of this territory, but it is a wider problem, of general type, because very often there is a lack of appropriate structures to run the business. Therefore, it could be suggested to constantly monitor territories under the transport and the environment point of view. By this way services can be supplied in an appropriate and efficient way. Only after, new infrastructures with optimal capacity can be created in accordance with the expressed and potential needs.

With particular regard to Gallipoli it is desirable the implementation of transport demand management tools through a regulation in terms of accessibility and of modal diversion in favour of environmental sustainable means. These needs have to be considered in the light of seasonal problems of congestion which affects the quality of life in the urban centre. In this case it can be useful drastic interventions in the summer period so to favour bicycles and public transport organised with a higher frequency and a geographic extension of the service also to the sea. Finally, for this city, the desired modal switch and rational behavioural transport choices could be easily reached through appropriate information supplied with highly accessible tools of ICT applied to all integrated transport modes.

Acknowledgments. The author thank the Municipality of Brindisi and the Management of URBANETS Project. Results presented in this paper are also part of the research project “Il processo decisionale nella scelta degli investimenti: aspetti economici, geografici, finanziari e traduttivi” supported by the University of Bari Aldo Moro.

References

[1] Green, W.H. (1997), Econometric analysis, Prentice-Hall International, London.

[2] Cascetta, E.-Papola, A. (2001), Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand, Transportation Research, part C, vol. 9, issue 4, pages 249-263.

[3] Maddala, G.S. (1999) Limited-dependent and qualitative variables in econometrics, Cambridge University Press, Cambridge.

Multi-resolution and spatial Independent Component Analysis approaches for geo-referred and time-varying mobile phone data

P. Zanini[1]

[1] *MOX – Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133, Milano (Italy); paolo.zanini@polimi.it*

Abstract. *The aim of this work is to provide different statistical tools to catch meaningful and useful information from geo-referred quantities varying along time. In particular a mobile-phone traffic dataset is analyzed to decompose the spatiotemporal information in order to identify spatial and temporal patterns. Two different approaches have been followed. The first one is an Independent Component Analysis (ICA) approach, where sources are assumed to be spatial stochastic processes on a lattice, in order to take into account the spatial dependence between pixels. This method is called spatial colored Independent Component Analysis (Shen, Truong, Zanini, 2014). The second one is a multi-resolution approach, where a temporal sparsity to the final representation is imposed through a wavelet-inspired data-driven procedure. This method is called Hierarchical Independent Component Analysis (Secchi, Vantini, Zanini, 2014). Results highlight urban features related to residential, leisure and mobility activities.*

Keywords. *Mobile-phone Data; Independent Component Analysis; Spatial Stochastic Processes; Multi-resolution Analysis*

1 Introduction

The aim of this work is to provide different statistical tools to catch meaningful and useful information from geo-referred quantities varying along time. This kind of data is present in a wide range of different applications.
In environmental analysis, for instance, the measure of pollution in a geo-referred area (e.g. a city, a river, etc...) at different instants of time need to be studied in order to face the environmental pollution problem. In meteorology, often, the analysis of temperatures or wind velocities in a geographic zone across time is considered for a lot of purposes. This work is focused on a case-study where the quantity of interest is a measure of mobile-phone traffic intensity evaluated on a rectangular grid over the city of Milan (Italy) across two weeks. The main purpose of the analysis is to decompose this spatiotemporal dataset in order to identify spatial and temporal patterns characterizing specific locations and/or specific periods. These patterns will be associated to different population behaviors related to residential, leisure and mobility activities. This can be useful, for instance, for urban planning and for real-time monitoring of the urban dynamics.

2 Data description and methods

The analyzed dataset is courtesy of a research agreement between Telecom Italia and the Politecnico di Milano. It consists in the evaluation of Erlang, a dimensionless intensity measure of the use of the mobile network, in a rectangular lattice L_0 of $n = 10573$ pixels covering the metropolitan area of Milan for a global surface of more than 700 km^2. Measurements are taken every 15 minutes for a period of two weeks. Data have been preprocessed exploiting a Fourier basis expansion, as described in [2]. Then, processed data consist in the Erlang measurements in the lattice L_0 at $p = 200$ instants of time regularly spaced in the time interval of one week. An example for a specific pixel and for a fixed instant of time is shown in Figure 1.

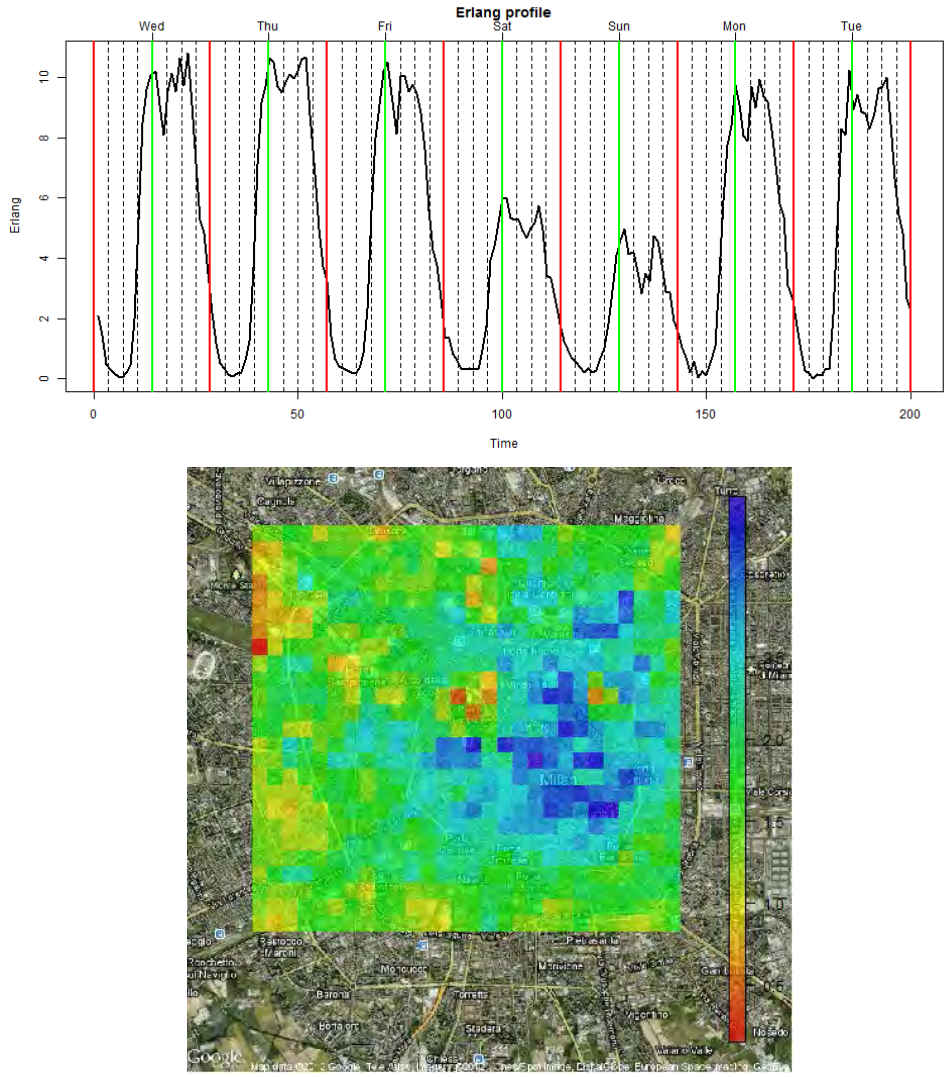


Figure 1: On the top: Erlang signal for a specific pixel. On the bottom: Erlang surface at a fixed instant of time.

Aim of the analysis is to decompose the observed signal as a time-varying linear combination of a reduced number, say K, of time-invariant source surfaces. Specifically, for a fixed pixel l_i and a fixed

time interval t_j:

$$x_{ij} = s_{i1}a_{j1} + ... + s_{iK}a_{jK},$$

where s_{ik} represents the contribution of the kth source in the pixel l_i and a_{jk} is the intensity of the kth source at the jth time interval. This problem fits in the Blind Source Separation framework

$$X = SA.$$

Indeed the purpose of the analysis is to represent the $n \times p$ data matrix X as the product of two matrices, the $p \times K$ basis (loadings) matrix A, and the $n \times K$ source (scores) matrix S, where A gathers the temporal profiles and S the spatial maps of the decomposition. Then, spatial maps can be associated to different urban features, while temporal profiles describe the activation periods of such features.
Two different approaches have been considered:

- an Independent Component Analysis (ICA [1]) approach, where sources are modeled as independent random variables. In fact, sources are assumed to be stochastic processes on a lattice, in order to take into account the spatial dependence between pixels. This analysis is performed through the algorithm named spatial colored Independent Component Analysis (scICA [4]), which works in the frequency domain. It exploits the Whittle likelihood and a kernel based nonparametric algorithm in order to estimate the spatial processes and their spectral densities;
- a multi-resolution approach, where the interest is in finding a sparse and multi-resolution estimate for the basis matrix A. The method used is named Hierarchical Independent Component Analysis (HICA [3]). It provides a multi-resolution (wavelet inspired) data-driven basis, through a hierarchical procedure with the application of ICA on pairs of variables at each step. In this way a temporal sparsity to the final representation is imposed.

The two approaches have been applied to the Telecom dataset, providing interesting results in terms of phenomenological interpretation. The comparison is done not with the purpose to establish which method is better, but to show the different features catched by the two approaches considered, and how these results can contribute to highlight patterns related to urban activities.

References

[1] Hyvarinen, A., Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. *Neural Networks* **13**, 411–430.

[2] Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., Vitelli, V. (2015). Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region. In Paganoni, A., Secchi, P. (eds.). *Advances in Complex Data Modeling and Computational Methods in Statistics Contributions to Statistics*, Springer, 133–147.

[3] Secchi, P., Vantini, S., Zanini, P. (2014). Hierarchical Independent Component Analysis: a multi-resolution non-orthogonal data-driven basis. *Mox Report 01/2014*, Dipartimento di Matematica, Politecnico di Milano.

[4] Shen, H., Truong, Y., Zanini, P. (2014). Independent Component Analysis for Spatial Stochastic Processes on a Lattice. *Mox Report 38/2014*, Dipartimento di Matematica, Politecnico di Milano.