# Addressing the change of support problem with non Gaussian distributions: an application to hourly rainfall prediction 

F. Bruno, D. Cocchi, F. Greco and E. Scardovi*

Department of Statistical Sciences "Paolo Fortunati", University of Bologna;
francesca.bruno@unibo.it,daniela.cocchi@unibo.it,fedele.greco@unibo.it, elena.scardovi2@unibo.it *Corresponding author


#### Abstract

We address the change of support problem in a non Gaussian model; in particular, we predict occurrence and accumulation of hourly rainfall in the Emilia-Romagna region by modeling the relationship between rain gauge and radar data. A basic formulation exploits radar information only at the grid cells containing rain gauges locations; an enrichment consists in exploiting neighbourhood information via a weighted mean of a latent spatial process for addressing spatial misalignment. A Bayesian hierarchical model is specified for each formulation, both sharing a zero-inflated likelihood, and with gamma distribution on the positive semiaxis. Results are assessed and compared in terms of point and probabilistic forecasts.


Keywords. Change of support; Data fusion; Hierarchical Bayesian model; Zero-inflated; Hourly rainfall

## 1 Introduction

The increasing availability of data sources induces the search of statistical tools for data assimilation. In particular, when measurements are provided on different spatial supports (e.g. point vs grid, or different resolutions), the change of support problem must be addressed. Two main alternative approaches are proposed in literature. Bayesian melding [4] combines observational data with computer model output via a latent point-level process driving both sources of data. On the other hand, downscalers use to accommodate for spatial misalignment through Bayesian hierarchical models relating the monitoring station data and the computer model output, using a spatial linear model with spatially varying coefficients (see for example [1]); in this case, the model is fitted only on monitoring sites, thus reducing the computational burden with respect to the first approach. These models are usually specified under Gaussian likelihood. Our contribution includes the downscaler in a model dealing with non Gaussian data. In particular, in this work we show an application to hourly rainfall data with zero inflated-gamma distribution.

## 2 Emilia-Romagna rainfall data description

Rainfall measurements are essential for public authorities, being the basis for hydrological models and risk monitoring; knowledge of rainfall amounts with high spatial resolution is useful for water resource planning and management. Direct measurements are provided by rain gauges in sparsely distributed locations. Radar data are available on fine-pixel grids, thus overcoming the problem of sparseness of the rain gauges network. However they consist in unreliable indirect measurements, despite accurate pre-processing performed by removing systematical and occasional biases. We focus on hourly rainfall data in the Italian Emilia-Romagna region. Radar data consist of hourly rainfall maps with 1 x 1 km grid cell resolution ( $\sim 48000$ pixels), within a circle of 125 km radius; in this area, about 300 rain gauges are available. The data are heterogeneous in space and time, and contain many zero values, corresponding to dry hours. Eight rain events, characterised by different duration and meteorological conditions, occurred in September-October 2010, are analysed. A description can be found in [2], where separate modeling of different hours is performed, suggesting a slight preference for the gamma distribution over the competing lognormal for modeling positive rain amounts.

## 3 Model specification for spatial prediction

A flexible choice for dealing with the remarkable quantity of zero measurements is proposed in [8] by the zero-inflated model, that constitutes the first level of the hierarchy:

$$
\begin{equation*}
p\left(Y_{s} \mid R, \pi_{s}\right)=\pi_{s} I_{Y_{s}=0}+\left(1-\pi_{s}\right) p\left(X_{s}\right) I_{Y_{s}>0} \quad s \in S_{G} \tag{1}
\end{equation*}
$$

where $\pi_{s}$ is the probability of zero and $Y_{s}$ is the rain gauge measurement at location $s, p\left(X_{s}\right)=p\left(Y_{s} \mid Y_{s}>0\right)$ and $S_{G}$ is the set of rain gauges locations. $R$ denotes radar information; it will assume different meaning according to the downscaler specification.
Further levels of the hierarchy specify the models for rainfall probability and amount:

- rain occurrence: $\operatorname{probit}\left(1-\pi_{s}\right)=\gamma_{1}+\gamma_{2} \log (R)+\varepsilon_{s} \quad$ where

$$
\begin{equation*}
\boldsymbol{\epsilon} \mid \sigma_{\varepsilon}^{2}, \phi_{\varepsilon} \sim M V N\left(\mathbf{0}, \sigma_{\varepsilon}^{2} \boldsymbol{\Sigma}_{\varepsilon}\right) \quad \text { and } \quad \boldsymbol{\Sigma}_{\varepsilon}\left(s, s^{\prime}\right)=\exp \left(-\phi_{\varepsilon} d_{s s^{\prime}}\right) \tag{2}
\end{equation*}
$$

- rain accumulation:

$$
\begin{gather*}
X_{s} \mid \mu_{s}, \tau \sim \operatorname{Gamma}\left(\tau, \tau / \mu_{s}\right) \quad \text { where } \quad \log \mu_{s}=\beta_{1}+\beta_{2} \log (R)+\alpha_{s}  \tag{4}\\
\boldsymbol{\alpha} \mid \sigma_{\alpha}^{2}, \phi_{\alpha} \sim \operatorname{MVN}\left(\mathbf{0}, \sigma_{\alpha}^{2} \boldsymbol{\Sigma}_{\alpha}\right) \quad \text { and } \quad \boldsymbol{\Sigma}_{\alpha}\left(s, s^{\prime}\right)=\exp \left(-\phi_{\alpha} d_{s s^{\prime}}\right) . \tag{5}
\end{gather*}
$$

Both rain occurrence and rain accumulation are modeled by exploiting radar measurements as a covariate; the spatial information is captured by normal random effects whose correlation decreases exponentially with the Euclidean distance $d$ between locations (according to (3) and (5)).
We denote with $R_{P}$ the radar value in pixel $P, P \in S_{R}$, where $S_{R}$ indicates the set of all the available pixels in the grid, $N_{R}=\# S_{R}=48047$. In [2] the change of support problem is addressed by relating $Y_{s}$ to the radar measurement $R_{P(s)}$, where $P(s)$ is the grid cell containing location $s$; this corresponds to replacing $R$ with $R_{P(s)}$ in equations (2) and (4). From now on this formalization will be denoted as Model 1. As an alternative, in order to face misalignment and to fruitfully exploit the whole radar map, following [1] radar information $R$ can be substituted by a stochastic weighted mean. In particular, since diagnostic tools (like Brier Score) showed high skills in predicting rainfall occurrence without the need for further
enrichments (see [2]), we now focus on modeling the rainfall amounts. More precisely, the second part of formula (4) is replaced by

$$
\begin{equation*}
\log \mu_{s}=\beta_{1}+\beta_{2} \log \left(R_{\overline{P s}}\right)+\alpha_{s} \tag{6}
\end{equation*}
$$

where $R_{\bar{P} s}$ is an $s$-specific weighted mean of radar values over $S_{R}$ :

$$
\begin{equation*}
R_{\bar{P} s}=\sum_{P \in S(R)} w_{P s} R_{P}, \quad s \in S_{G} \tag{7}
\end{equation*}
$$

This model will be denoted as Model 2 in the following. The weights $w$ are stochastic, relying on a unique latent $N_{R}$-dimensional Gaussian process $Q$ with exponential covariance function, defined over the grid; the influence of each component $Q_{P}$ of such a process on a specific location $s \in S_{G}$ is smoothed according to the distance between $s$ and the centroid $c_{P}$ of pixel $P$ using an exponential kernel $K$ :

$$
\begin{equation*}
w_{P s}=\frac{K\left(s-c_{P}\right) \exp \left(Q_{P}\right)}{\sum_{P^{\prime} \in S(R)} K\left(s-c_{P^{\prime}}\right) \exp \left(Q_{P^{\prime}}\right)} . \tag{8}
\end{equation*}
$$

The decay parameters for $K$ and for the covariance function of $Q$ are exogenously fixed in order to maintain the influence of the pixels within a range of 5 km .
In order to reduce the computational burden, a predictive process $Q^{*}$ is actually estimated instead of $Q$, defined on a rougher grid ( 1 pixel every 16 in both directions) and analytical properties of the multivariate normal distribution are exploited for recovering the whole process. Noninformative hyperpriors are proposed for hyperparameters. Parameter estimation is performed through Markov chain Monte Carlo algorithms, implementing a Gibbs Sampler with Metropolis-Hastings steps.

## 4 Assessment and comparison between two model specifications

We compare the two alternative models in a random set of 50 validation sites for each hour. The Bayesian approach provides whole predictive distributions, allowing any further inspection like the assessment of uncertainty or of the probability of exceeding certain thresholds. Calibration, i.e. consistency between the forecasts and the empirical distribution, is confirmed by non-randomized PIT histograms (see [3]) which are both nearly uniform as desired; PIT histogram for Model 1 is reported, as an example, in the left-hand panel of Figure 1. Calibration and sharpness are simultaneously quantified via continuous rank probability score (CRPS). CRPS quantile decomposition plot (see [7]) reported in the central panel of Figure 1 shows that Model 2 performs better than Model 1 presenting lower values with respect to all quantile levels, with greater differences for central values.


Figure 1: Non-randomized PIT histograms for Model 1, CRPS quantile decomposition plot, and scatterplot of the means of the predictive distributions obtained under the two models.

The scatterplot of the means of the predictive distributions obtained under the two models, reported in the right-hand panel of Figure 1, shows the ability of Model 2 to predict higher rainfall quantities. Point predictions from the two models are evaluated and ranked according to the mean squared error (MSE) and the mean absolute error (MAE), which are consistent with respect to the mean and median respectively; information about raw radar information is also provided. Synthetic results for each rainfall event are reported in Table 1; both models provide lower values of the chosen scoring functions with respect to raw radar, confirming that calibration is successful. The enrichment proposed in Model 2 about the way of addressing the change of support problem improves both probabilistic and point forecasts.

| Event | CRPS |  | MSE |  |  | MAE |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mod 1 | Mod 2 | Mod 1 | Mod 2 | Radar | Mod 1 | Mod 2 | Radar |
| E1 | 0.25 | 0.25 | 0.34 | 0.39 | 2.57 | 0.33 | 0.31 | 0.94 |
| E2 | 0.67 | 0.66 | 5.40 | 5.20 | 8.57 | 0.90 | 0.88 | 1.40 |
| E3 | 0.42 | 0.41 | 2.05 | 1.96 | 5.06 | 0.58 | 0.56 | 1.38 |
| E4 | 0.57 | 0.52 | 4.27 | 3.76 | 13.40 | 0.79 | 0.69 | 1.86 |
| E5 | 0.69 | 0.66 | 7.12 | 6.59 | 11.95 | 0.95 | 0.86 | 1.54 |
| E6 | 0.18 | 0.18 | 0.53 | 0.53 | 1.48 | 0.23 | 0.23 | 0.68 |
| E7 | 0.35 | 0.35 | 1.03 | 1.08 | 4.13 | 0.48 | 0.46 | 0.91 |
| E8 | 0.17 | 0.18 | 0.35 | 0.34 | 1.49 | 0.23 | 0.23 | 0.63 |

Table 1: Assessment of probabilistic forecasts via CRPS and of point forecasts via MSE and MAE.

Acknowledgments. We wish to thank ARPA-SIMC Emilia Romagna for introducing us to the problem, providing data and performing data pre-processing, and Tilmann Gneiting for the joint work on the evaluation of predictions. This research work was funded by an ARPA-SIMC / Department of Statistical Sciences agreement (nr. 20/2012) and by a FIRB 2012 grant (project no. RBFR12URQJ; title: Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions) for research projects by the Italian Ministry of Education, Universities and Research.

## References

[1] Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2011) Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality. Biometrics 68, 837-848.
[2] Bruno, F., Cocchi, D., Greco, F. and Scardovi, E. (2013) Spatial reconstruction of rainfall fields from rain gauge and radar data. Stoch. Environ. Res. Risk Assess. 28 1235-1245.
[3] Bruno, F., Greco, F. and Scardovi, E. (2014) Assessment of Bayesian models for rainfall field reconstruction. Proceedings of the $47^{\text {th }}$ Scientific Meeting of the Italian Statistical Society, Cagliari, Italy, June 11/13.
[4] Fuentes, M. and Raftery, A. E. (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61, 36-45.
[5] Gneiting, T. (2011) Making and evaluating point forecasts. J. Amer. Stat. Assoc. 106, 746-762.
[6] Gneiting, T. and Raftery, A. (2007) Strictly proper scoring rules, prediction, and estimation. J. Amer. Stat. Assoc. 102, 359-378.
[7] Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold and quantile weighted proper scoring rules. Journal of Business and Economic Statistics, 29, 411-422.
[8] Sloughter, J., Raftery, A.E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian Model Averaging. Mon. Weath. Rev. 135, 3209-3220.

