On moderation in a Bayesian log-contrast compositional model with a total. Interaction between extreme temperatures and pollutants on mortality

Germà Coenders[1,2], Javier Palarea-Albadalejo[3], Marc Saez[1,2], Maria A. Barceló[1,2]

1 Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain (germa.coenders@udg.edu, marc.saez@udg.edu,. antonia.barcelo@udg.edu)
2 CIBER of Epidemiology abnd Public health (CIBERESP), Madrid, Spain
3 Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Spain (Javier.palarea@udg.edu)

Background and Objectives

Compositional data describe parts of a whole (e.g., pollutant concentrations, economic indicators) and require specialized analysis since standard statistical methods often yield misleading results. Log-ratio transformations (e.g., alr, ilr) properly handle their relative nature by mapping compositions to real space while preserving scale invariance. However, many applications like air pollution studies require modeling both relative (compositional) and absolute (total) effects, which traditional approaches conflate. The *T*-space theory provides a framework to coherently separate and analyze these dual aspects.

We extend compositional regression to incorporate moderation effects (interactions), enabling investigation of how variables influence each other's impact on outcomes. A novel parameterization maintains interpretability through elasticities while respecting the zero-sum constraints essential for compositional data analysis. This approach uniquely integrates these constraints with Bayesian inference via INLA, offering computational efficiency alongside theoretical rigor.

The method's flexibility supports both frequentist and Bayesian frameworks, making it widely applicable across environmental science, econometrics, and other fields where compositional data analysis is crucial. By simultaneously addressing relative compositions, total magnitudes, and their interactions, this work provides a more complete toolkit for modern compositional data modeling.

Methods

We used a time-series ecological design for the summer months (June to September, both inclusive) of 2022. All the variables were analysed at the (contextual) level of 379 health areas (ABS is the acronym in Catalan: *Àrea Bàsica de Salut*) in which the territory of Catalonia is divided.

Outcome variable was the daily deaths from all causes for all ages and was obtained from the Spanish National Statistics Institute (INE) at the census tract level, for the total population of Catalonia. They were aggregated at the ecological level for each ABS. We obtained information on the semi-hourly levels of the meteorological variables (maximum and minimum temperatures and relative humidity) for 2012–2022 from the 189 automatic meteorological stations in the Network of Automatic Meteorological Stations (excluding 12 meteorological stations that are at an altitude of 1,500 metres or more) of the Meteorological Service of Catalonia (METEOCAT), which were active in the period (open data). We averaged the semi-hourly data to obtain the daily data. Data on the hourly levels of the air pollutants: airborne particulate matter with a diameter of 10 microns ($\mu g/m^3$) or less (PM_{10}), nitrogen dioxide (NO_2) , ozone (O_3), sulphur dioxide (SO_2) and carbon monoxide (CO) were obtained from the 95 automatic monitoring stations of the XVPCA network (Catalan Network for Pollution Control and Prevention) (open data). As with meteorological data, we averaged the hourly data (the 8-hour moving average in the case of O_3) to obtain the daily data. *Extreme of maximum temperature*: Indicator that the maximum temperature exceeds the trigger temperature on a given day (the 95th percentile of the frequency distribution of the daily values of the spatiotemporal prediction of the maximum temperature in the period 2012-2021, excluding the very possible outliers of 2015 and 2022).

Let us consider a non-closed composition **x** comprising *D* strictly positive pollutant concentrations, with the purpose of modeling the association between **x** and extreme temteraures, in the role of explanatory or predictor variables, and a real-valued variable y in the role of dependent or response. Here, the relationship with the explanatory terms is established through the (natural) logarithm link function as the linear predictor[1]:

$$
\begin{aligned}
\log(\theta'_{it}) = {} & \beta_0 + \beta_1 \log(PM_{10,i,t-4}) + \beta_2 \log(NO_{2,i,t-4}) + \beta_3 \log(O_{3,i,t-4}) + \beta_4 \log(CO_{i,t-4}) \\
& + \beta_5 \log(SO_{2,i,t-4}) + \beta^{(z)} \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \beta_1^{(z)} \log(PM_{10,i,t-4}) \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \beta_2^{(z)} \log(NO_{2,i,t-4}) \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \beta_3^{(z)} \log(O_{3,i,t-4}) \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \beta_4^{(z)} \log(CO_{i,t-4}) \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \beta_5^{(z)} \log(SO_{2,i,t-4}) \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \beta^{(t)} \text{total}_{it-4} + \beta^{(t)} \text{total}_{it-4} \text{extreme\_maximum\_temperature}_{i,t-4} \\
& + \eta_i + S(ABS_i) + \tau s_{it} + offset(\log(population_i))
\end{aligned}
$$

$$
with \sum_{j=1}^{5} \beta_j = 0 \ and \ \sum_{j=1}^{5} \beta_j^{(z)} = 0
$$

where θ'_{it} was the conditional risk of a death in ABS *i* on day *t*; extreme heat denoted an indicator of maximum temperature extreme (with lag l=4); η_i was the unstructured random effect capturing individual heterogeneity; τs_{it}, the structured random effects, capturing annual seasonality; S() the structured random effects, normally distributed with zero mean and a Matérn covariance function that depends on the distance between ABS, controlling spatial dependence; and Population denoted the population in ABS *i*.

Note that:

i) We included a **soft constraint**[2,3], $\sum_{j=1}^{5} \beta_j = 0 \ and \ \sum_{j=1}^{5} \beta_j^{(i)} = 0$, so that in Bayesian inference the prior distribution did not interfere with the log-ratio representation.
ii) All variables **lagged by 4 days** relative to the response variable. This was set after testing several lags and achieving the lowest WAIC for a lag equal to 4.
iii) We considered, in addition to the potential effect of relative changes among parts of the composition, the **total magnitude of the composition** (total), capturing the contribution of their actual absolute values to the variation in the response variable[4,5].
iv) The sum ($\beta_j + \beta^{(t)}$) is directly interpretable as an **elasticity**. Thus, a small relative increase (e.g. by 1%) in x_j, while keeping the remaining parts unaltered, leads to an increase of ($\beta_j + \beta^{(t)}$)% in the response *y* (when extreme_maximum_temperature= 0).
v) This elasticity can be decomposed via the *T*-sapace. When all parts of the composition increase simultaneously by 1/*D*%, then the response *y* increases by $\beta^{(t)}$% while holding extreme_maximum_temperature = 0. When the ratio of x_j to any other part increases by 1%, while keeping the total constant, the response *y* increases by β_j%, with extreme_maximum_temperature = 0.
vi) As to the **moderation or interaction effects**, the coefficients $\beta_j^{(i)}$ and $\beta^{(t,i)}$ represent the change in the compositional and total effects for the case of extreme_maximum_temperature = 1. If all the parts increase simultaneously by 1/*D*%, then the response y increases by ($\beta^{(t)} + \beta^{(t,i)}$)%. If the ratio of x_j to any other part increases by 1%, the response y increases by ($\beta_j + \beta_j^{(i)}$)%.

Inferences were made following a Bayesian perspective, using the INLA approach[6,7]. We used priors that penalize complexity (called PC priors). These priors are robust in the sense that they do not have an impact on the results, and, in addition, they have an epidemiological interpretation[8].

The implementation of the soft constraint in INLA was done through a new function called **A.local.** The A.local function allows users to impose local constraints on the linear predictor. Previously, if a linear predictor contained more than one element of the same model component, it required a global redefinition through complicated design matrices, which also involved irrelevant parts of the model. With A.local, this is avoided in most cases by efficiently applying constraints directly to specific subsets of the model predictors.

Results

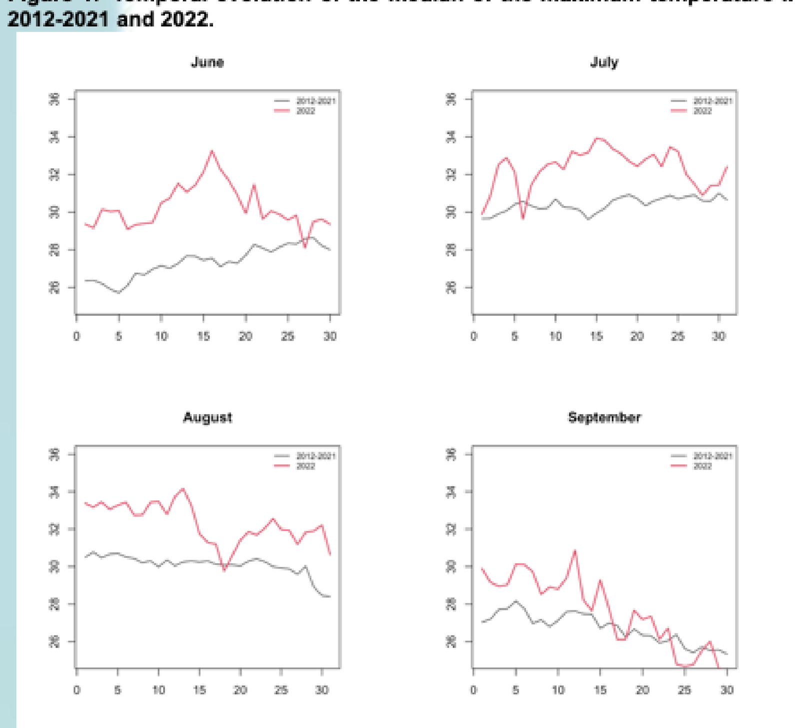
Table 1 Results from the INLA estimation of the specified zero-inflated negative binomial generalised mixed model

	Mean[1]	Prob.[2]
β_1	-0.002	0.623
β_2	0.010	0.914
β_3	-0.001	0.558
β_4	-0.006	0.856
β_5	-0.001	0.561
$\beta^{(t)}$	0.006	0.983
$\beta^{(z)}$	0.071	>0.99
$\beta_1^{(i)}$	0.025	0.937
$\beta_2^{(i)}$	0.005	0.613
$\beta_3^{(i)}$	-0.011	0.766
$\beta_4^{(i)}$	0.003	0.580
$\beta_5^{(i)}$	-0.022	0.922
$\beta^{(t,i)}$	0.002	0.629

[1]Posterior means of the relevant model coefficients.
[2]Posterior probabilities of being positive or negative.

- The coefficient β_2 has a posterior probability of being positive equal to 0.914. Then, if the ratio of the concentration of NO_2 to any other pollutant increases by 1%, while keeping the total pollution constant, the expected number of deaths four days later will increase by 0.010%. Note that this may equally apply to days of extremely high temperature or days of normal temperature, given that the corresponding interaction coefficient $\beta_2^{(i)}$ has a low posterior probability.
- The coefficient $\beta^{(t)}$ has a posterior probability equal to 0.983 of being positive. Then, if the total pollution increases by 1/D = 1/5%, while the relative composition remains unaltered, the expected number of deaths four days later will increase by 0.006%. If we scale the results for an easier interpretation, when the total pollution increases by 1%, the expected number of deaths four days later will increase by 0.006x5 = 0.030%. Note that this may equally be applied for days of extremely high temperature and days of normal temperature, since the corresponding interaction coefficient $\beta^{(t,i)}$ has a low posterior probability of 0.629.
- The coefficient $\beta^{(z)}$ is estimated to have a posterior probability greater than 0.999 of being positive. This indicates that the expected number of deaths will increase by 7.4% (e0.071 = 1.074) four days after the occurrence of an extreme temperature event when pollutant concentrations are on their average.
- The coefficient $\beta_1^{(i)}$ has a posterior probability of being positive equal to 0.937. Then, if the ratio of the concentration of PM_{10} to any other pollutant increases by 1%, while keeping the total pollution constant, the expected number of deaths will increase by 0.025%. Note that this may only hold during days of extremely high temperature, since the corresponding main effect coefficient β_1 has a low posterior probability of 0.623. Moreover, the negative sign and high absolute value of $\beta_5^{(i)}$ (posterior probability of being negative equal to 0.922) indicates that increased mortality is expected mainly when PM_{10} is traded for SO_2.

Figure 1.- Temporal evolution of the median of the maximum temperature in 2012-2021 and 2022.



References

1.- Coenders G, Palarea-Albadalejo J, Saez M, Barceló MA. Moderation effects and elasticities in compositional regression with a total. Application to Bayesian spatiotemporal modelling of all-cause mortality from environmental stressors. *arXiv: 2505.07800v1*. doi: 10.48550/arXiv.2505.07800.
2.- Aitchison J, Bacon-Shone J. Log contrast models for experiments with mixtures. *Biometrika* 1984; 71(2): 323–330 (1984). doi: 10.2307/2336249.
3.- Zhang L, Zhang X, Yi N. Bayesian compositional generalized linear models for analyzing microbiome data. *Statistics in Medicine* 2024; 43(1): 141–155. doi: 10.1002/sim.9946
4.- Coenders G, Martín-Fernández JA, Ferrer-Rosell B. When relative and absolute information matter: Compositional predictor with a total in generalized linear models. *Statistical Modelling* 2017; 17(6):494-512. doi: 10.1177/1471082X17710398.
5.- Mota A, Saez M, Coenders G. Compositional and Bayesian inference analysis of the concentrations of air pollutants in Catalonia, Spain. *Environmental Research* 2022; 204:112388. doi: 10.1016/j.envres.2021.112388.
6.- Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Socisty Series B Statistical Methodology* 2009; 71:319-392. doi:j.1467-9868.2008.00700.x.
7.- Rue H, Riebler A, Sørbye H, Illian JB, Simpson DP, Lindgren FK. 2017. Bayesian computing with INLA: A review. *Annual Reviews of Statistics and its Applications* 2017; 4: 395-421. doi: annurev-statistics-060116-054045.
8.- Simpson DP, Rue H, Martins TG, Riebler A, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science* 2017; 32(1):1-46. doi: 10.1214/16-STS576.

This research was supported by the Spanish Ministry of Science and Innovation and ERDF-a way of making Europe [grant number PID2021-123833OB-I00], the Spanish Ministry of Health [grant number CIBERCB06/02/1002], the Department of Research and Universities of Generalitat de Catalunya [grant numbers 2021SGR01197 and 2023-CLIMA-00037] and AGAUR and the Department of Climate Action, Food and Rural Agenda of Generalitat de Catalunya [grant number 2023-CLIMA-00037].