Supplementary data of the SALMANTICOR study

Spatial analysis

We will combine multiple factor analysis (MFA) and Cokriging statistics procedures to provide a spatial analysis of the SALMANTICOR population.

Our study will inquire and analyzed N individuals from M municipalities. Q questionnaires were handed to all the participants. Let X_{nmq} be a matrix block where n is the number of participant of a m municipality and k is the correspondent questionnaire of our departing matrix D_{MxQ}.

Therefore, depending on the type of k questionnaire, we will employ a PCA, MCA or CA, to each block X_{nmq} obtaining $\overline{Y}_{mq} = \frac{1}{\lambda_{mq}} Y_{mq}$ where λ_{mq} is its first singular value.

Hence, we join all the resulting $\overline{Y}_{mq}$ forming a $\overline{X}_{MxF}$ matrix where M are the municipalities and F the resulting factors.

$$\overline{X}_{mf} = [\overline{Y}_{m1}|\overline{Y}_{m2}| \ldots |\overline{Y}_{mf}| \ldots |\overline{Y}_{mF}]$$

Finally, a generalized PCA is applied on $\overline{X}_{MxF}$

After performing MFA we will proceed to project the resulting coordinates that represents our municipalities over the resulting L latent variables obtaining R_{MxL}.

Adding the spatial coordinates u to each municipality we attain $Z(u) = [u|R]$. Once we get the $Z(u)$ matrix, we will apply a spatial interpolator such as Cokriging.

We will then describe the spatial behavior of our samples using variograms. Variograms are illustrations of how the semivariance acts in function of the distance. Semivariance is defined as half the expectation between two different values at two locations (u and $u + h$), and is used in univariate analyses. To transfer our analysis to a multivariate problem we will need to build crossvariograms.

A crossvariogram γ_{ij} describes the degree of spatial dependence of our projected variables measuring the variation between two samples depending on the distance h (also known as lag) between them.

After this step, we will define

$$\Gamma(h) = \frac{1}{2}\left[\left(Z_i(u) - Z_i(u + h)\right) \cdot \left(Z_j(u) - Z_j(u + h)\right)\right]$$

with $i, j = 1 \ldots M$ and hence, the crossvariogram

Using a more practical approach, we will need to build a set of experimental crossvariograms based on our matrix $Z(u)$.

Therefore, we will obtaine $\frac{L(L+1)}{2}$ experimental semivariograms, and subsequently these direct and crossvariograms will need to be fitted. The different parts of a theoretical semivariogram are:

Nugget: It represents variability at small distances ($h \approx 0$).

Sill: The semivariance b value at which the semivariogram levels off.

Range: The a distance at which the semivariogram reaches the sill value.

The Linear Model of Coregionalization (LMC) permits all the $\frac{L(L+1)}{2}$ semivariograms to be fitted as linear combinations of S basic semivariogram functions (Gaussian, Exponential, Spherical, etc). The LMC can be expressed as a multivariate nested semivariogram model.

$$\Gamma(h) = \sum_{s=1}^{S} B_s g_s(h)$$

where $\Gamma(h)$ is the S×S matrix of semivariogram values at lag h, and B_s is the S×S matrix of sills of the basic semivariogram function $g_s(h)$. B_s has to be positive semidefinite, to assure that the variance-covariance matrix is also positive.

Once $\Gamma(h)$ is set, we will need to interpolate over the different polygons that represents the municipalities and shape the province of Salamanca. For fulfilling this task, we will apply Cokriging.

Cokriging is the multivariate extension of kriging, whose main purpose is to compute a weighted average of the sample values in close proximity to a grid point, polygon or volume. It searches for the best linear unbiased estimator, based on assumptions on covariances. There are different procedures such as ordinary, universal, or simple Cokriging.

As an example, we present simple Cokriging.

$$\bar{Z}_{i_0}(u_0) = m_{i_0} + \sum_{i=1}^{L} \sum_{\alpha=1}^{M} w_\alpha^i (Z_i(u_\alpha) - m_i)$$

where u_0 is an unsampled municipality and u_α a sample location, w_α^i is the weight and m corresponds to the means of our variables. We can associate a simple cokriging system to this estimator as $C_{ij} w_i = c_{ii_0}$, where C_{ij} is the M×M covariance matrix, and c_{ii_0} is the $M_0 \times M$ covariance matrix between the unsampled and sample locations.

Machine learning

The following table describes the selected machine learning (ML) algorithms to be used in the SALMANTICOR study.

Algorithm	Type	Description
Random Forest	Combine methods	Classification ensemble through a combination set of non-correlated independently decision trees
Gradient Boosting	Combine methods	Ensemble technique in which decision trees are not independently, but sequentially

Algorithm	Type	Description
Logistic regression	Regression	The go-to method for categorical or binary classification
K-nearest Neighbors	Supervised classification	Classifies each unlabeled example by the majority label among its k-nearest neighbors in the training set
Support Vector Machine	Supervised classification	Classification and regression technique through construction of separating hyperplanes to maximize the margin and to produce a generalization ability
Linear discriminant analysis	Linear discriminant	Searches for directions in the data that have the largest variance and subsequently project the data onto it combining Fisher vectors
Naive Bayes classifier	Probabilistic supervised classification	The Bayesian classification is used as a probabilistic learning method