Machine learning to deal with uncertainty in knowledge base for multivariate clustering applied to spatial analysis

Stephane Bourrelly[1]

Antonino Marvuglia[2]

Ian Vázquez-Rowe[3]

[1]University of Lyon 3 – UMR 7300 (ESPACE), France
[2]Luxembourg Institute of Science and Technology (LIST), Luxembourg
[3]Pontificia Universidad Católica del Perú, Peru

*Corresponding author: s.bourrelly@hotmail.fr

Abstract
We present a Tailor Made Machine Learning (TMML) methodology combining different clustering algorithms, spatial statistical methods and cartographic tools. The methodology is currently being programmed in an R package, especially designed to handle multivariate spatial datasets. We highlight the strengths of unsupervised clustering for the management of environmental health phenomena, also pointing out several uncertainty sources affecting the results of our analysis. In particular, we acknowledge that the traditional hierarchical clustering is usually applied without performing dynamic reallocations, integrating spatial key-concepts or discussing the quality of outputs. Therefore we describe the foundations of the TMML methodology, which is applied to deal with these uncertainties, as well as with the variety of possible outputs. The R functions are applied to the spatial dataset included in the package so to illustrate the procedure to apply for identifying the most accurate clustering output, in the context of a sustainable agriculture example in Luxembourg.

Keywords
Clustering, lattice, uncertainty, decision-making, agriculture.

I INTRODUCTION

When dealing with multidimensional datasets, clustering algorithms are often used to produce a reduced knowledge-base to simplify the representation of phenomena. The aim of multivariate clustering applied to spatial contexts is to form groups as heterogeneous as possible, composed by spatial units as similar possible. When clustering performs on lattice; spatial typologies are created in order to classify the administrative areas in meaningful clusters and summarise several dimensions of phenomena. In this way knowledge can be extracted from the multidimensional complexity with the purpose to conceive local policies, e.g. to ensure the economic growth and prevent socio-ecological conflicts (Delgado and Romero, 2016). The Hierarchical Clustering Analysis (HCA) is the most applied unsupervised method by public stakeholders, e.g. for the sustainable management of the agricultural activities (Boyacioglu and Boyacioglu 2008).

Unsupervised classification methods provide objective representations, unlike supervised methods which constrain the results by a known response variable. Unfortunately, even though

HCA presents a few advantages, the results are not optimal. Therefore some dynamic reallocation methods have been conceived to overcome this drawback (Hennig and Liao, 2010). However Levine (2000) notes that they are scarcely used in social sciences as well as the quality criteria which are rarely presented to discuss the statistical significance and the meaning of typologies. Moreover Levine points out that unsupervised clustering is usually used without integrating the key concepts of the scope of application; e.g. when algorithms are applied without taking into account the interconnectivity of the spatial units. Obviously, these shortfalls contribute to increase results' uncertainty, which might give rise to controversial decisions.

In this paper we briefly describe the methodology of a so-called Tailor-Made Machine Learning (TMML), which combines several clustering and data-mining algorithms with spatial statistic operators as well as cartographic tools. The TMML provides a methodology for taking into account the interconnectivity of spatial units. We present some of the functions of the TMML R package currently under development, to compare and map the outputs of the implemented clustering. Functions are applied to the case study shapefile of the TMML package, representing the environmental concentrations of chemical substances generated by the application of agricultural fertilizers in the administrative areas in Luxembourg.

II MATERIAL

The shapefile of the TMML package was derived from a cadastre spatial layer, representing the agricultural land use on the administrative areas (communes) in the Grand-duchy. This polygon layer describes the 108.328 georeferenced parcels contained in the cadastre records in 2009. These agricultural surfaces are differentiated from their main Agricultural Land Class (ALC) among the 23 official types, such as cereals, oilseed, legumes, etc. (Figure 1).

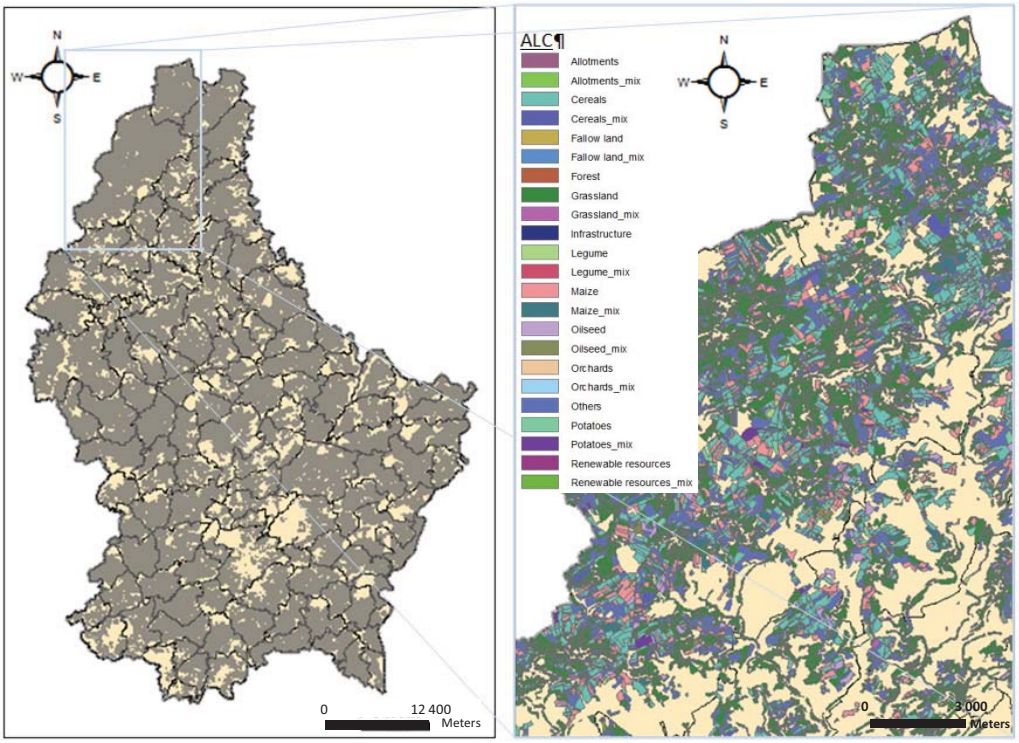

Figure 1: Luxembourg maps with 108.328 parcels named according to their main ALC.

In order to know the agricultural harvests in 2009 we used the agricultural statistics of the national database[1]. However, national statistics differentiate the agricultural surfaces from 27

[1] http://www.statistiques.public.lu/stat/ReportFolders/

different categories of crops; which have been matched according to the ALC description of cadastral data.

Fertilizers dosages (and in some cases yields data) have been derived from experts' communications, as well as from KTBL (2006). Chemical emissions from fertilizers application have been estimated according to (Nemecek and Kägi 2007). Finally the ArcGIS software has been used for aggregating, at the scale of communes, all the environmental emissions and computing the spatial indicators. We note x_k^j,the concentration level of the chemical substance 'j' in the administrative area of the commune 'k' for an environmental compartment, i.e. air, ground water or surface water (Table 1).

	AIR			GROUNDWATER		SURFACE WATER	
j	x_NH_3	x_NO_x	x_N_2O	x_P	x_PO4^{3-}	$x\_NO_3^-$	x_PO4^{3-}
substance	Ammoniac	Nitrogen oxide	Nitrous oxide	Phosphorus	Phosphate	Nitrate	Phosphate

Table 1: description of spatial indicators $x_.^j$.

The command

- shapeLux = data(ml.chemichal)

-loads polygon shapefile of the TMML package. It describes the 116 communes of Luxembourg in 2009 (names, surfaces, index) and provides the values of the 7 x_k^j. The Figure 2 displays the spatial distributions of the concentrations of $\mathbf{NO_x}$ in air, $\mathbf{NO_3^-}$ in surface water and $\mathbf{PO4^{3-}}$ in groundwater, expressed in kg per km² of administrative area.

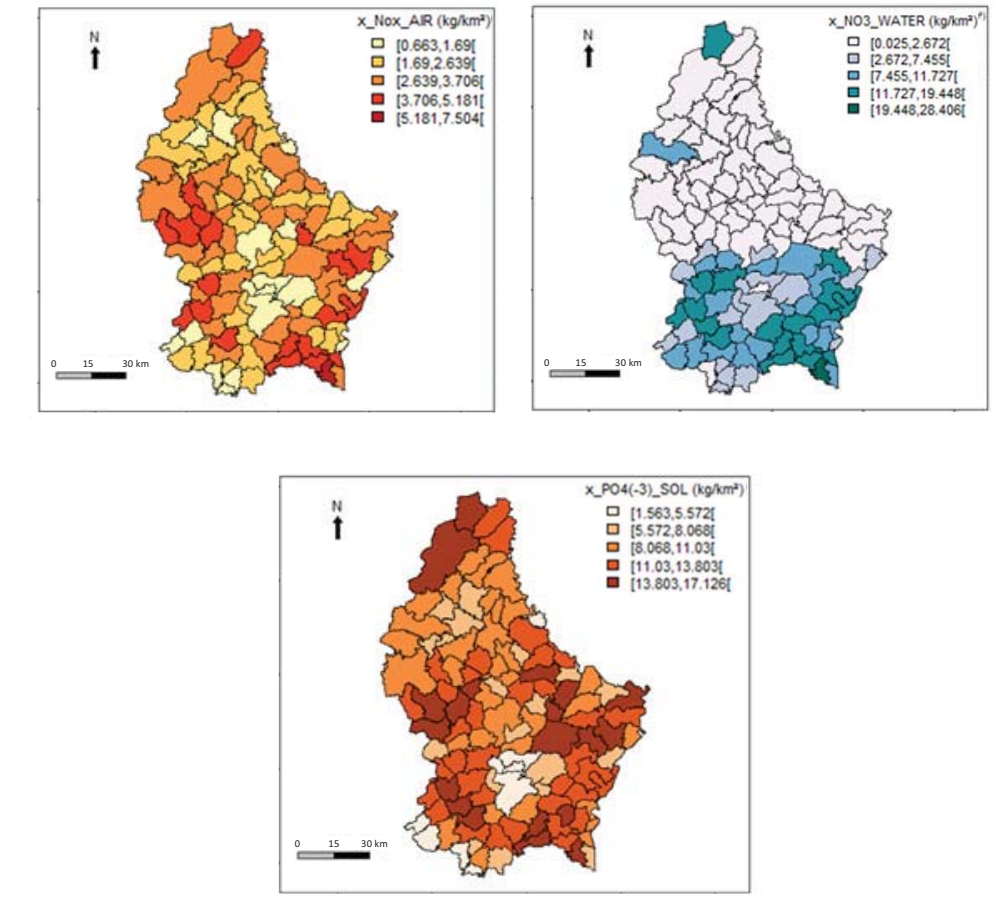

Figure 2: maps of $x_k^{NO_x}$ (upper left), $x_k^{NO_3^-}$ (upper right) and $x_k^{PO_4^{3-}}$ (lower).

III METHOD

The HCA algorithm performs an unsupervised classification for creating objective typologies. At the first step it groups the two most similar individuals (spatial units) so to form the first cluster (group). Then it iteratively maximises the inter-clusters inertia to group the other spatial units into nested cluster. Nested clusters are organized as a hierarchical tree called dendrogram, which allows specifying a number of clusters C.However with hierarchical clustering methods the composition of each of the generic clusters C_i is not optimal, due to the nested merging process. To overcome this drawback dynamic reallocation methods have been conceived in the literature for iteratively interchanging the cluster composition and minimising the intra-clusters inertia. The most popular algorithms are the Partitioning Around Medoïds (PAM) and K-means (Kaufmann and Rousseeuw, 1990).

Since the model hypotheses are different they provide different typologies. In addition they not allow objectively choosing a number of clusters; that is why they are usually initialised with the HCA outputs. The interpretation of typologies is a controversial issue, as only their statistical significance can be discussed from several quality criteria such as the R^2, the C-INDEX, the rate of inertia or the average silhouette (Hennig and Liao, 2010).

Moreover these clustering methods assume that individuals are independent, but each spatial unit influences the other ones located in its neighbourhood. Here we propose a methodology to integrate the spatial connectivity in the clustering analysis.

Firstly, in the TMML guidelines a conceptual neighbourhood pattern among those defined by Cliff and Ord (1981) should be chosen (Figure 3). In this way a neighbourhood matrix W is created. W contains weights w_{kl}, where $w_{kl} = 1$ if the spatial unit 'k' is directly adjacent to the spatial unit 'l' and $w_{kl} = 0$ otherwise. In order to test the interdependency between the adjacent values x_k^j, x_l^j the Moran's I statistic can be used (Gaetan and Guyon, 2010).

The tests can be replicated at several orders h of contiguity, i.e. for $h > 1$. In this case the neighbourhood matrix is termed $W(h)$ and the strength of spatial dependencies can be read on a correlogram. The correlogram chart displays the values of h as a function of the Moran's statistics $I(h)$ and their two-side confidence interval values, to denote the presence of a spatial autocorrelation (Gaetan and Guyon, 2010). In this way clustering algorithms can be applied to lagged values of x_k^j, noted xs_k^j to take into account the spatial dependency concept.

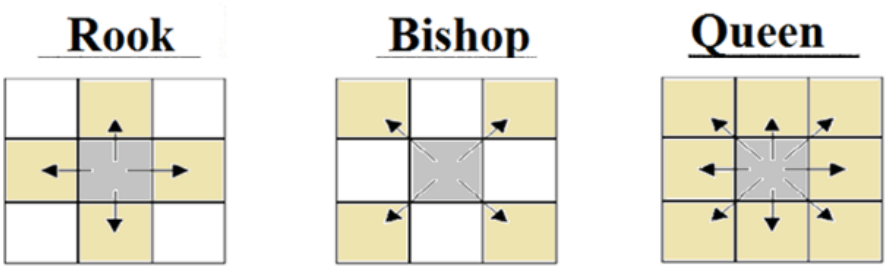

Figure 3: conceptual neighbourhood patterns at the order h=1.

IV RESULTS

The function

- ml.setC(x=X, update.C=4)

-performs a HCA and provides the chart of inertia differences between-clusters in decreasing orders, so to help in setting the number of groups C (Figure.4).

The function

- ml.clust.(shape=shapeLux, x=X, C=4, maps=T, bar.charts=T, ordinal=T, labCi = c("WEAK","MIDDLE","HIGH","MAJOR"))

-performs the HCA, K-means and PAM for merging the spatial units in C groups. It maps the distributions of class labels for the clustering outputs. With an ordinal typology the class labels 'labCi' are assigned according to the average centres of C_i When bar.charts=T bar charts are returned to summarise the characteristics of communes within the cluster C_i. The horizontal bars give differences in number of standard deviations, between the means of C_i and the overall average values for each chemical emission (Figure.4).

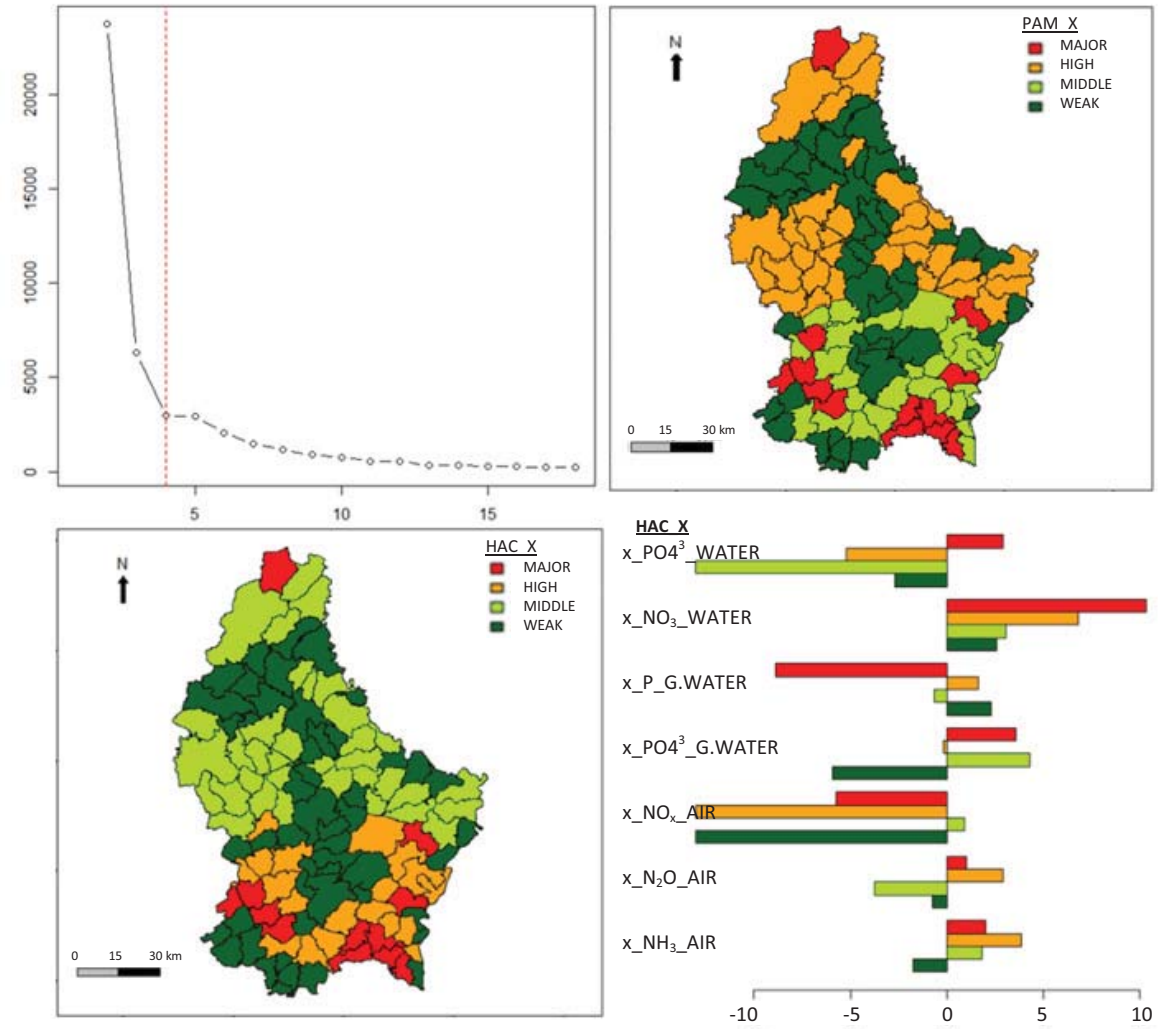

Figure 4: chart of inertia differences (upper left); maps of typologies for PAM (upper right) and HAC (lower left) with related bar chart (lower right).

The function

- ml.pattern(shape=shapeLux, x=X, j.lab="x_NOx", type="QUEEN", map=T Moran.test=T, correlogram="Moran", h.max=7)

-explores, for each spatial indicator j.lab, the interdependencies of values from a particular type of spatial pattern. The global autocorrelation of x_k^j is forecasted through the Moran test and the lag order of spatial dependencies, such as $h \leq$ h.max, can be assessed from the correlogram (Figure.5).

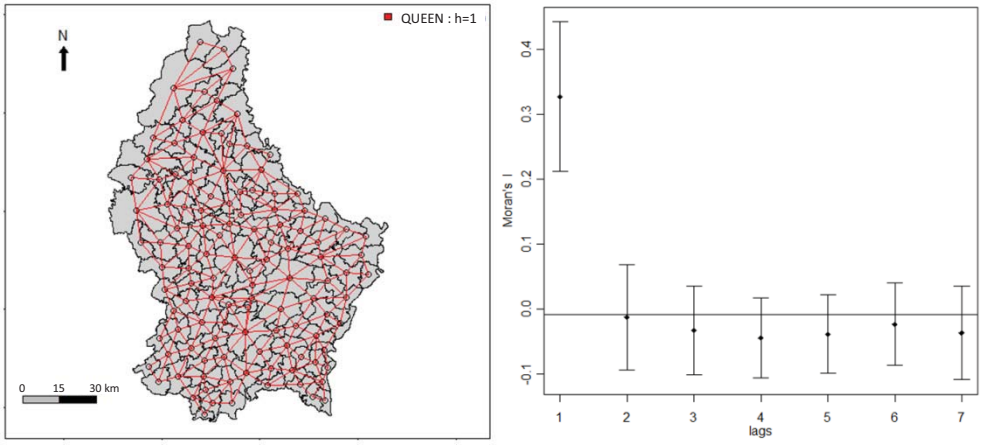

Figure 4: map of "Queen" connectivity pattern at the first order and Morans'I correlogram.

The function

- ml.clust.(…, …, spatial.patterns=rep("QUEEN", 7), h=c(1,1,1,2,2,3,5), mat.weight="W", quality=T, maps.xs=T)

-performs the HCA, K-means and PAM on the spatially lagged values of indicators $xs_{.}^{j}$, according to the spatial patterns, j-order dependencies h and standardised neighbourhood matrix, when mat.weight="W" (Figure.5). When quality=TRUE the quality criteria of clustering, highlighted in the method section, are returned.

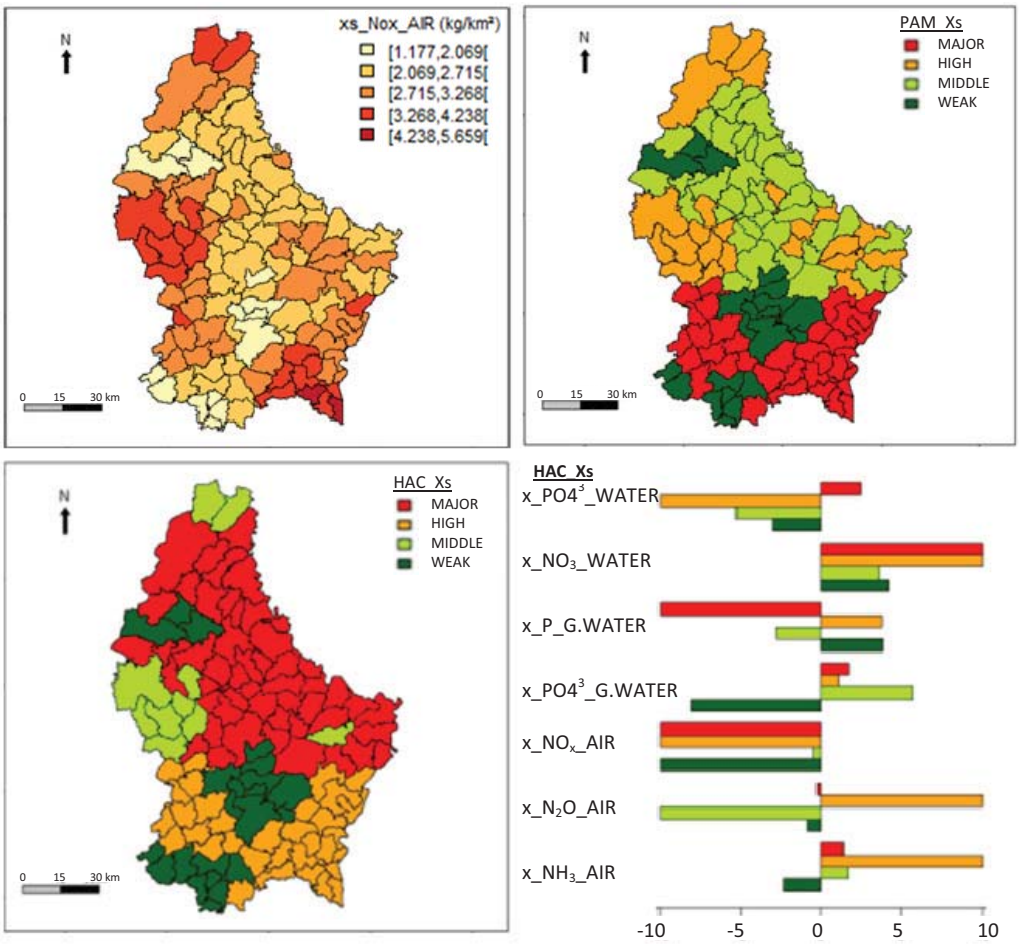

Figure 5: maps of xs_k^{Nox} and the typologies for PAM and HAC, with related bar chart.

V DISCUSSION

The variety of spatial typologies returned from the different clustering algorithms applied to the original or lagged spatial indicators, highlights the strong uncertainty induced in decision-making when the quality of outputs is not questioned.

Even though the uncertainties might be explored through some quality criteria, from the analyst's perspective it is important knowing to what extent the dynamic reallocations and the integration of spatial patterns really improve the statistical significance of the results.

In this respect an important uncertainty source in unsupervised clustering methods lays into the fact that the best clustering is usually chosen through statistical criteria. Indeed, the model selection issue is currently a recognised challenge in the unsupervised classification field (Hennig and Liao, 2010).

Ideally, the best clustering representation should be identified through a trade-off between the statistical significance of typologies and their spatial correlations with a decision criterion, identified by stakeholders. In particular, we point out the importance of defining the most suitable spatial typology for reducing the impacts of chemical emissions on eco-systems and therefore the cumulative socio-ecological inequalities. In future work, these latter could finally be estimated in line with the index of Morello-Frosch et al. (2011).

References

Boyacioglu H., Boyacioglu H. (2008). Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey. Environ Geol, 54, 275–282.

Cliff A., Ord J. (1981). The problem of spatial autocorrelation. In Scott A. London: Pion. 25-55.

Delgado A., Romero I. (2016). Environmental conflict analysis using an integrated grey clustering and entropy-weight method: A case study of a mining project in Peru. Environmental Modelling & Software, 77, 108-121.

Hennig C., Liao F. (2010). Comparing latent class and dissimilarity based. Department of Statistical Science, UCL, Department of Sociology. University of Illinois.

Gaetan C., Guyon X. (2010). Spatial Statistics and Modeling. Berlin: Springer.

Kaufman L., Rousseeuw P. (1990). Finding groups in data: an introduction to cluster analysis. New York: Wiley.

KTBL - Kuratorium für Technik und Bauwesen in der Landwirtschaft (Ed.) (2006). Faustzahlen für die Landwirtschaft, Darmstadt, Germany (in German).

Levine J. (2000). But What Have You Done For Us Lately. 29(1), 35-40

Morello-Frosch R, Zuk M, Jerrett M, Shamasunder B, Kyle AD. (2011). Understanding he cumulative impacts of inequalities in environmental health: implications for policy. *Health Aff (Millwood)* 30, 879–87.

Nemecek T., Kägi T. (2007). Life Cycle Inventories of Swiss and European Agricultural Production Systems. Final report ecoinvent V2.0 No. 15a. Agroscope Reckenholz-Taenikon Research Station ART, Swiss Centre for Life Cycle Inventories, Zurich and Dübendorf, CH, retrieved from: www.ecoinvent.ch.