Comparison of Dispersion Models by Using Fuzzy Similarity Relations

Angelo Ciaramella[1], Angelo Riccio[1], Stefano Galmarini[2],
Giulio Giunta[1], and Slawomir Potempski[3]

[1] Department of Applied Science,
University of Naples "Parthenope", Isola C4, Centro Direzionale, I-80143,
Napoli (NA), Italy
{angelo.ciaramella,angelo.riccio,giulio.giunta}@uniparthenope.it
[2] European Commission, DG Joint Research Centre,
Institute for Environment and Sustainability,
21020 Ispra (VA), Italy
stefano.galmarini@jrc.ec.europa.eu
[3] Institute of Atomic Energy, Otwock-Swierk, Poland
slawek@cyf.gov.pl

Abstract. Aim of this work is to introduce a methodology, based on the combination of multiple temporal hierarchical agglomerations, for model comparisons in a multi-model ensemble context. We take advantage of a mechanism in which hierarchical agglomerations can easily combined by using a transitive consensus matrix. The hierarchical agglomerations make use of fuzzy similarity relations based on a generalized Łukasiewicz structure. The methodology is adopted to analyze data from a multi-model air quality ensemble system. The models are operational long-range transport and dispersion models used for the real-time simulation of pollutant dispersion or the accidental release of radioactive nuclides in the atmosphere. We apply the described methodology to agglomerate and to individuate the models that characterize the predicted atmospheric pollutants from the ETEX-1 experiment.

Keywords: Fuzzy Similarity, Hierarchical Agglomeration, Ensemble Models, Air Pollutant Dispersion.

1 Introduction

Clustering is an exploratory tool in data analysis that arises in many different fields such as data mining, image processing, machine learning, and bioinformatics. One of the most popular and interesting clustering approaches is the hierarchical agglomerative clustering. In this work we introduce a novel methodology based on fuzzy similarity relations that permits to combine multiple temporal hierarchical agglomerations.

This methodology has been applied to data concerning the real-time forecasting of atmospheric compounds from the *ENSEMBLE* system [5,6,7]. ENSEMBLE is a web-based system aiming at assisting the analysis of multi-model data

R. Pirrone and F. Sorbello (Eds.): AI*IA 2011, LNAI 6934, pp. 57–67, 2011.

provided by many national meteorological services and environmental protection agencies worldwide for the real-time forecasting of deliberate/accidental releases of harmful radionuclides (e.g. Fukushima, Chernobyl).

In previous works [15,12] an approach for the statistical analysis of multimodel ensemble results has been presented. The authors used a well-known statistical approach to multimodel data analysis, i.e., *Bayesian Model Averaging*, which is a standard method for combining predictive distributions from different sources. Moreover, similarities and differences between models were explored by means of correlation analysis. In [13] the authors investigate some basic properties of multi-model ensemble systems, which can be deduced from general characteristics of statistical distributions of the ensemble membership. Cluster-based approaches [1,2,3] have also been developed and applied. These approaches discriminate between data that are less dependent (in the statistical sense), so that "redundant" information can be more easily discarded and equivalent performance can be achieved with a considerable lower number of models.

In this paper we generalize these clustering approaches, by introducing a new methodology based on fuzzy similarity relations that allows to combine multiple hierarchical agglomerations, each for a different forecasting leading time.

We conjecture that this framework is amenable to easily incorporate observations that may become available during the course of the event, so as to improve the forecast by "projecting" observations onto the hierarchical combination of clusters.

The paper is organized as follows. In Sections 2 and 3 some fundamental concepts on t-norms and fuzzy similarity relations are given. The proposed methodology is detailed in Section 3.3. Finally, in Section 4 some experimental results obtained by applying this methodology on an ensemble of prediction models are described. Conclusions and future remarks are given in Section 5.

2 Norms and Residuum

In this Section we introduce some basic terminologies and successively we outline the minimum requirements a fuzzy relation should satisfies in order to correspond to a dendrogram and in what cases dendrograms can be aggregated into a consensus matrix.

The popularity of *fuzzy logic* comes mainly from many applications, where linguistic variables are suitably transformed in fuzzy sets, combined via the conjunction and disjunction operations by using continuous triangle norms or co-norms, respectively. Moreover, it offers the possibility of soft clustering, in contrast with algorithms that output hard (crisp or non-fuzzy) clustering of data.

A fundamental concept in fuzzy logic is that of *norm* [9]. A *triangular norm* (*t-norm* for short) is a binary operation t on the unit interval $[0,1]$, i.e., a function $t : [0,1]^2 \to [0,1]$, such that for all $x, y, z \in [0,1]$ the following four axioms are satisfied:

$$\begin{array}{llr} t(x,y) = t(y,x) & & (\textit{commutativity}) \\ t(x,t(y,z)) = t(t(x,y),z) & & (\textit{associativity}) \\ t(x,y) \leq t(x,z) & \text{whenever } y \leq z & (\textit{monotonicity}) \\ t(x,1) = x & & (\textit{boundary condition}) \end{array} \quad (1)$$

Several parametric and non-parametric t-norms have been introduced [9] and recently a their generalized version has been studied [4]. The four basic t-norms are $t_{\mathbf{M}}$, $t_{\mathbf{P}}$, $t_{\mathbf{L}}$ and $t_{\mathbf{D}}$ given by, respectively:

$$\begin{array}{lclr} t_{\mathbf{M}}(x,y) & = & \min(x,y) & (\textit{minimum}) \\ t_{\mathbf{P}}(x,y) & = & x \cdot y & (\textit{product}) \\ t_{\mathbf{L}}(x,y) & = & \max(x+y-1,0) & (\textit{Łukasiewicz t-norm}) \\ t_{\mathbf{D}}(x,y) & = & \begin{cases} 0 & \text{if } (x,y) \in [0,1]^2 \\ \\ \min(x,y) & \text{otherwise} \end{cases} & (\textit{drastic product}) \end{array} \quad (2)$$

In the following, we concentrate on the $t_{\mathbf{L}}$ norm.

The union and intersection of two unit interval valued fuzzy sets are essentially lattice operations. In many applications, however, lattice structure alone is not rich enough to model fuzzy phenomena. An important concept is the residuated lattice and its structure appears, in one form or in another, in practically all fuzzy inference systems, in the theory of fuzzy relations and fuzzy logic. One important operator here is the *residuum* $\rightarrow_t$ defined as

$$x \rightarrow_t y = \bigvee \{z | t(z,x) \leq y\} \quad (3)$$

where $\bigvee$ is the *union* operator and for the left-continuous basic t-norm $t_{\mathbf{L}}$ is given by

$$x \rightarrow_{\mathbf{L}} y = \min(1-x+y,1) \ (\textit{Łukasiewicz implication}) \quad (4)$$

Moreover, let p a fixed natural number in a *generalized Łukasiewicz structure*, we have

$$\begin{array}{rl} t_{\mathbf{L}}(x,y) = & \sqrt[p]{\max(x^p+y^p-1,0)} \\ \\ x \rightarrow_{\mathbf{L}} y = & \min(\sqrt[p]{1-x^p+y^p},1) \end{array} \quad (5)$$

Finally, we define as *bi-residuum* on a residuated lattice the operation

$$x \leftrightarrow_t y = (x \rightarrow_t y) \wedge (y \rightarrow_t x) \quad (6)$$

where $\wedge$ is the *meet*.

For the left-continuous basic t-norm $t_{\mathbf{L}}$ we have

$$x \leftrightarrow_{\mathbf{L}} y = 1 - \max(x,y) + \min(x,y) \quad (7)$$

3 Fuzzy Similarity

A binary *fuzzy relation* R on $U \times V$ is a fuzzy set on $U \times V$ ($R \subseteq U \times V$). *Similarity* is a fuzzy relation $S \subseteq U \times U$ such that, for each $u, v, w \in U$

$$\begin{array}{lll} S\langle u,u\rangle & = \quad 1 & (\textit{everthing is similar to itself}) \\ S\langle u,v\rangle & = S\langle v,u\rangle & (\textit{symmetric}) \\ t(S\langle u,v\rangle, S\langle v,w\rangle) & \leq S\langle u,w\rangle & (\textit{weakly transitive}) \end{array} \tag{8}$$

It is also well known that a fuzzy set with membership function $\mu : X \to [0,1]$ generate a fuzzy similarity S defined as $S\langle a,b\rangle = \mu(a) \leftrightarrow_t \mu(b)$ for all $a, b \in X$. We also note that, let $t_{\mathbf{L}}$ be the Łukasiewicz product, we have that S is a fuzzy equivalence relation on X with respect to $t_{\mathbf{L}}$ iif $1 - S$ is a *pseudo-metric* on X. Further, a main result is the following [17,16]:

Proposition 1. *Consider n Łukasiewicz valued fuzzy similarities S_i, $i = 1, \ldots, n$ on a set X. Then*

$$S\langle x,y\rangle = \frac{1}{n}\sum_{i=1}^{n} S_i\langle x,y\rangle \tag{9}$$

is a Łukasiewicz valued fuzzy similarity on X.

3.1 Min-transitive Closure

If a similarity relation is *min-transitive* (t = min in (8)), it is called a *fuzzy-equivalence relation.* Each fuzzy-equivalence relation can be graphically described by a *dendrogram* [10]. Therefore, the requirement for the existence of the dendrogram, for a similarity matrix, is the transitivity.

The methodology introduced in this paper uses a min-transitive closure [11]. The transitive closure is obtained by computing a sufficiently high power of the given similarity matrix. Let n the dimension of a relation matrix, the transitive closure R^T of R is calculated by

$$R^T = \bigcup_{i=1}^{n-1} R^i \tag{10}$$

where R^{i+1} is defined as

$$R^{i+1} = R^i \circ R \tag{11}$$

The composition $R \circ S$ of fuzzy relations R and S is a fuzzy relation defined by

$$R \circ S\langle x,y\rangle = \mathrm{Sup}_{z\in X}\{R\langle x,z\rangle \wedge S\langle z,y\rangle\} \tag{12}$$

$\forall x, y \in X$ and where $\wedge$ stands for a t-norm (e.g., min operator) [11]. Using this methodology the min-transitive closure R^T can be computed by the algorithm described in Algorithm 1.

Algorithm 1. Min-transitive closure

```
1: Input $R$ the input relation
2: Output $R^T$the output transitive relation
3: 1. Calculate $R^* = R \cup (R \circ R)$
   2. if $R^* \neq R$ replace $R$ with $R^*$ and go to step 1
   else $R^T = R^*$ and the algorithm terminates.
```

The transitive property of binary relations is closely related to the theory of the graphs. In other words, if a relation is represented as a directed graph, then computation of transitive closure of this relation is equivalent to finding the tightest path between each pair of vertices. The strength of a path is determined by the minimum of the weights on that path.

3.2 Dendrogram Description Matrices

As previously described, any dendrogram could be associated with a fuzzy equivalence relation and, equivalently, with its matrix representation if the min-transitive closure property is satisfied. The elements of a fuzzy equivalence matrix describe the similarity between objects. Moreover we have that [11]

Lemma 1. *Letting R be a similarity relation with the elements $R\langle x, y\rangle \in [0, 1]$ and letting D be a dissimilarity relation, which is obtained from R by*

$$D(x, y) = 1 - R\langle x, y\rangle \tag{13}$$

then D is ultrametric iif R is min-transitive.

There is a one-to-one correspondence between min-transitive similarity matrices and dendrogram. The correspondence between ultrametric dissimilarity matrices and dendrograms is also on-to-one. In other words, a dendrogram could be generated corresponding to a dissimilarity matrix if it is *ultrametric.*

3.3 Agglomerative Methodology

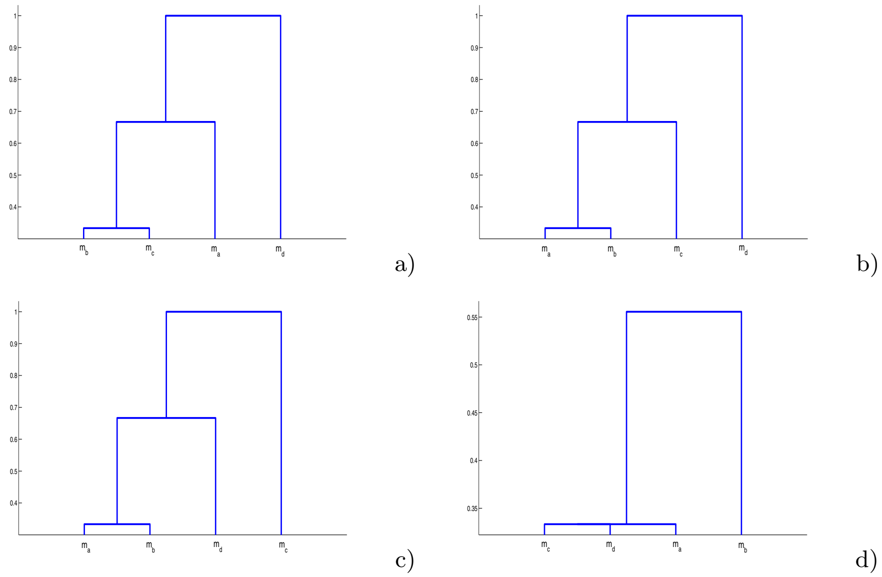
We remark that the aim is to agglomerate, by an unsupervised methodology, the distributions obtained by the ensemble models at different times. Substantially a hierarchical tree (dendrogram), that permits to cluster models that have similar behavior, must be obtained. We calculate the similarity (or dissimilarity) matrix between the distributions of the models by using the fuzzy similarity described in equation 9. Successively the algorithm described in Algorithm 1 is applied to obtain the min-transitive closure.

We also may express the information by *fuzzy set.* A simple way is to describe the *membership functions* by the following equation [17]

$$\mu(\mathbf{x}_i) = \frac{\mathbf{x}_i - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)} \tag{14}$$

Algorithm 2. Combination of dendrograms

```
1: Input $S^{(i)}$, $1 \leq i \leq L$ $L$ input similarity matrices (dendrograms)
2: Output $S$ the resulted similarity matrix (dendrogram)
   1. Aggregate the similarity matrices to a final similarity matrix $S = Aggregate(S^{(1)}, S^{(2)}, \ldots, S^{(L)})$
   a. Let $S^*$ be the identity matrix
   b. For each $S^{(i)}$ calculate e $S^* = S^* \cup (S^* \circ S^{(i)})$
   c. c. If $S^*$ is not changed $S = S^*$ and goto step 3 else goto step 1.b
3: Create the final dendrogram from the $S$
```



a)

b)

c)

d)

Fig. 1. Combination algorithm: a-b-c) input dendrograms; d) combined hierarchy

where $\mathbf{x}_i = [x_1^i, x_2^i, \ldots, x_L^i]$ is the i-th observation vector of the L models.

Successively, we apply the agglomerative hierarchical clustering approach to obtain the dendrogram. A consensus matrix that it is representative of all dendrograms is obtained by combining the transitive closure and equation 12 (i.e., max-min) [11]. The algorithm to obtain the final dendrogram is described in Algorithm 2.

In Figure 1 we show a realistic agglomeration result. In Figures 1a-b-c three input hierarchies to be combined are plotted. Four models are considered, namely m_a, m_b, m_c and m_d, respectively. In Figure 1d we show the final result obtained calculating the dendrogram on the similarity matrix. The result seems to be rational, because the output hierarchy contains the clusters (m_a, m_b, m_c) and (m_a, m_b, m_c, m_d) at different levels, and each of these clusters are repeated at least in two out of the three input dendrograms [11].

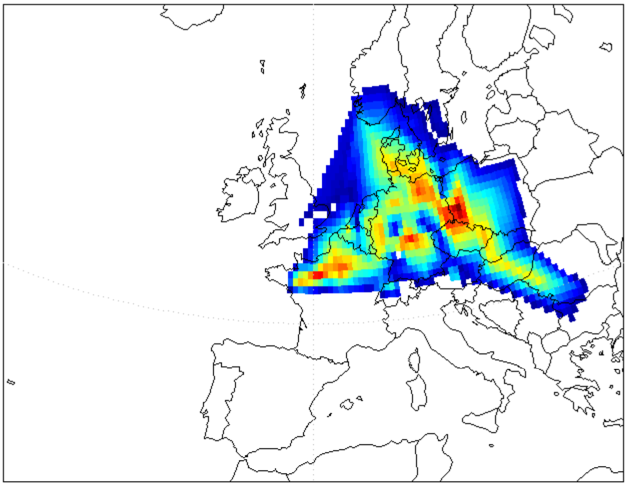


Fig. 2. ETEX-1 integrated (in time) observations

4 Experimental Results

In this Section we propose the results obtained applying the described methodology to compare mathematical operational long-range transport and dispersion models used for the real-time simulation of pollutant dispersion.

In [7,15] the authors analyzed the output of multi-model ensemble results for the ETEX-1 experiment. They already showed that the "*Median Model*" provided a more accurate reproduction of the concentration trend and estimate of the cloud persistence at sampling locations.

The ETEX-1 [8] experiment concerned the release of pseudo-radioactive material on 23 October 1994 at 16:00 UTC from Monterfil, southeast of Rennes (France). Briefly, a steady westerly flow of unstable air masses was present over central Europe. Such conditions persisted for the 90 h that followed the release with frequent precipitation events over the advection area and a slow movement toward the North Sea region. In Figure 2 we show the integrated concentration after 78 hours from release. Several independent groups worldwide tried to forecast these observations. The ensemble is composed by 25 members. Each simulation, and therefore each ensemble member, is produced with different atmospheric dispersion models and is based on weather fields generated by (most of the time) different *Global Circulation Models* (GCM). All the simulations relate to the same release conditions. For details on the groups involved in the exercise and the model characteristics, refer to [7] and [8].

Now we describe the phases needed to analyze this dataset. The first step is the *fuzzification*. Namely, equation 14 is used on the estimated model concentrations at each time level. Successively, a similarity matrix (dendrogram) is obtained for the concentrations at different times (by using equation 9, Łukasiwicz

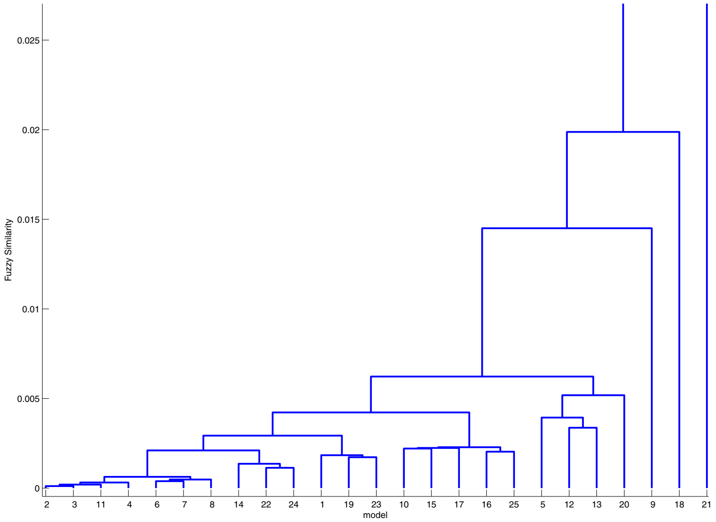


Fig. 3. Combined fuzzy similarity dendrogram

with $p = 1$). Finally the representative similarity matrix is estimated making use of Algorithm 2. A particular of the dendrogram obtained on the integrated concentrations after 78 hours is plotted in Figure 3. In this figure the information on the abscissa are related to the models and those on the ordinate are related to the model data similarities obtained by using the fuzzy similarity. As an example, in Figure 4 we show some distributions of the models. The distributions in Figures 4a-b are very close and this is confirmed by the dendrogram. Instead the model in Figure 4c has a diffusive distribution far from the other distributions and also confirmed by the dendrogram.

The hierarchical mechanism permits to clusterize the observations in a fixed number of clusters. A *Mean Square Error* (MSE) between each model and the median value of the cluster where it belongs is determined. For each cluster the model with the minimum MSE is considered. Finally the *median model* of these selected models is calculated and it is compared with the real observation by using the RMSE. Moreover, varying the number of clusters the models that have the best approximation of the real observation can be defined (see [15] and [3] for more details). In Figure 5 we show the *Root Mean Square Error (RMSE)* obtained varying the number of clusters. In this case the best approximation is obtained by using 6 clusters.

As can be inferred from the analysis of this figure, a lower RMSE does not necessarily corresponds to the use of a large number of models; similar (or even better) performance can be achieved with a few models; even more interestingly, since the selection framework is not based on the prior knowledge of experimental values, the satisfactory comparison of selected subset of models with

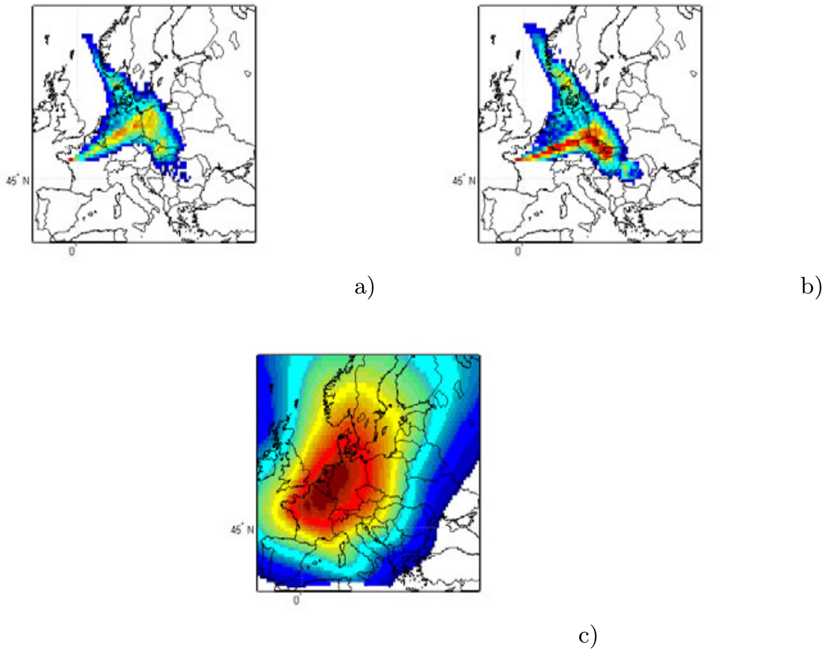


Fig. 4. Model distributions: a) model 22; b) model 24; model 21

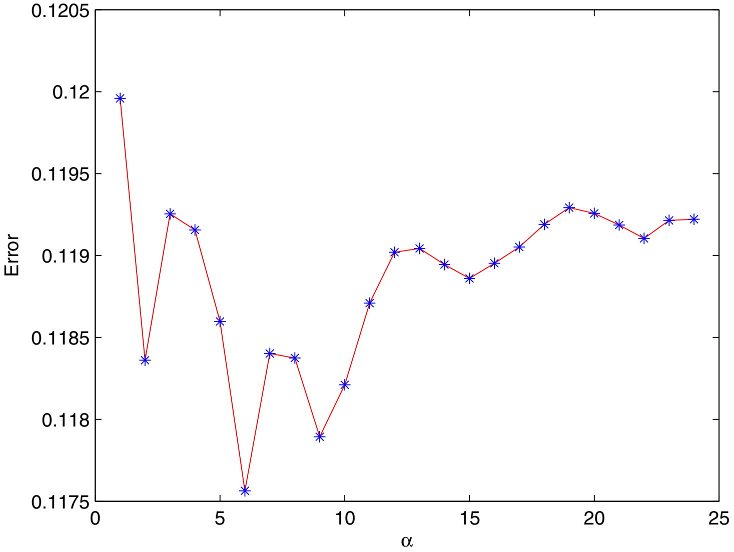


Fig. 5. RMSE varying the clustering number

experimental values suggest promising perspectives for the systematic reduction of ensemble data complexity. Furthermore, comparing this new methodology with the previous one, by using the consensus matrix we add temporal information that permits to obtain a more robust and realistic analysis.

5 Conclusions

In this work we introduced a methodology, based on the combination of multiple temporal hierarchical agglomerations, for model comparison in a multi-model ensemble context. Here we suggest to use fuzzy similarity relations in a Łukasiewicz structure. We remark that further studies can be made by using also different fuzzy similarities (e.g., [14]). Moreover, we take advantage of a mechanism in which hierarchical agglomerations can be easily combined by using a transitive consensus matrix. The proposed methodology is able to combine multiple temporal hierarchical agglomerations of dispersion models used for the real-time simulation of pollutant dispersions. The results show that this methodology is able to discard redundant temporal information and equivalent performance can be achieved considering a lower number of models reducing, the data complexity. In the next future, further studies could be conducted on real pollutant dispersions (e.g., Fukushima) and on the structure utilized in the fuzzy similarity relations.

References

1. Ciaramella, A., Cocozza, S., Iorio, F., Miele, G., Napolitano, F., Pinelli, M., Raiconi, G., Tagliaferri, R.: Interactive data analysis and clustering of genomic data. Neural Networks 21(2-3), 368–378 (2008)
2. Napolitano, F., Raiconi, G., Tagliaferri, R., Ciaramella, A., Staiano, A., Miele, G.: Clustering and visualization approaches for human cell cycle gene expression data analysis. International Journal of Approximate Reasoning 47(1), 70–84 (2008)
3. Ciaramella, A., Giunta, G., Riccio, A., Galmarini, S.: Independent Data Model Selection for Ensemble Dispersion Forecasting. In: Okun, O., Valentini, G. (eds.) Applications of Supervised and Unsupervised Ensemble Methods. SCI, vol. 245, pp. 213–231. Springer, Heidelberg (2009)
4. Ciaramella, A., Pedrycz, W., Tagliaferri, R.: The Genetic Development of Ordinal Sums. Fuzzy Sets and Systems 151, 303–325 (2005)
5. Galmarini, S., Bianconi, R., Bellasio, R., Graziani, G.: Forecasting consequences of accidental releases from ensemble dispersion modelling. J. Environ. Radioactiv. 57, 203–219 (2001)
6. Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J.C., Bellasio, R., Bompay, F., Buckley, R., Bouzom, M., Champion, H., D'Amours, R., Davakis, E., Eleveld, H., Geertsema, G.T., Glaab, H., Kollax, M., Ilvonen, M., Manning, A., Pechinger, U., Persson, C., Polreich, E., Potemski, S., Prodanova, M., Saltbones, J., Slaper, H., Sofief, M.A., Syrakov, D., Sorensen, J.H., Van der Auwera, L., Valkama, I., Zelazny, R.: Ensemble dispersion forecasting–Part I: concept, approach and indicators. Atmos. Environ. 38, 4607–4617 (2004a)

7. Galmarini, S., Bianconi, R., Addis, R., Andronopoulos, S., Astrup, P., Bartzis, J.C., Bellasio, R., Buckley, R., Champion, H., Chino, M., D'Amours, R., Davakis, E., Eleveld, H., Glaab, H., Manning, A., Mikkelsen, T., Pechinger, U., Polreich, E., Prodanova, M., Slaper, H., Syrakov, D., Terada, H., Van der Auwera, L.: Ensemble dispersion forecasting–Part II: application and evaluation. Atmos. Environ. 38, 4619–4632 (2004b)
8. Girardi, F., Graziani, G., van Veltzen, D., Galmarini, S., Mosca, S., Bianconi, R., Bellasio, R., Klug, W.: The ETEX project. EUR Report 181-43 EN. Office for official publications of the European Communities, Luxembourg, p. 108 (1998)
9. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Kluwer Academic Publishers, Dordrecht (2001)
10. Meyer, H.D., Naessens, H., Baets, B.D.: Algorithms for computing the min-transitive closure and associated partition tree of a symmetric fuzzy relation. Eur. Journal Oper. Res. 155(1), 226–238 (2004)
11. Mirzaei, A., Rahmati, M.: A novel Hierarchical-Clustering-Combination Scheme based on Fuzzy-Similarity Relations. IEEE Transaction on Fuzzy Systems 18(1), 27–39 (2010)
12. Potempski, S., Galmarini, S., Riccio, A., Giunta, G.: Bayesian model averaging for emergency response atmospheric dispersion multimodel ensembles: Is it really better? How many data are needed? Are the weights portable? Journal of Geophysical Research 115 (2010), doi:10.1029/2010JD014210
13. Potempski, S., Galmarini, S.: Est modus in rebus: analytical properties of multi-model ensembles. Atmospheric Chemistry and Physics 9(24), 9471–9489 (2009)
14. Rezaei, H., Emoto, M., Mukaidono, M.: New Similarity Measure Between Two Fuzzy Sets. Journal of Advanced Computational Intelligence and Intelligent Informatics 10(6) (2006)
15. Riccio, A., Giunta, G., Galmarini, S.: Seeking for the rational basis of the median model: the optimal combination of multi-model ensemble results. Atmos. Chem. Phys. 7, 6085–6098 (2007)
16. Sessa, S., Tagliaferri, R., Longo, G., Ciaramella, A., Staiano, A.: Fuzzy Similarities in Stars/Galaxies Classification. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 494–4962 (2003)
17. Turunen, E.: Mathematics Behind Fuzzy Logic. In: Advances in Soft Computing. Springer, Heidelberg (1999)