J2.6 A SPATIAL DATA MINING APPROACH FOR VERIFICATION AND UNDERSTANDING OF ENSEMBLE PRECIPITATION FORECASTING

Xuechao Yu*[1,2], Ming Xue[2,3], Limei Yang[4] and Le Gruenwald[4]

[1]NOAA/NWS/WDTB Cooperative Institute for Mesoscale Meteorological Studies
[2]School of Meteorology
[3]Center for Analysis and Prediction of Storms
[4]School of Computer Science,
University of Oklahoma, Norman, OK 73019

1. INTRODUCTION

As ensemble forecasting has been operational successfully for about 10 years for the global forecasting systems at centers such as NCEP and ECMWF, increasing emphasis is placed on mesoscale ensemble forecasting in recent years [e.g., the Storm and Mesoscale Ensemble Experiment (SAMEX) of 1998 (Hou et al., 2001)]. Since ensemble prediction provides an estimate of the forecast probability distribution of atmosphere states and generates huge amount of data, one of the main problems for its operational use is the design of manageable products for potential users. The common products include the ensemble and cluster means, standard deviations or spread, and probabilities of different events (Toth et al., 1997). One of the most widely used products is the "spaghetti" diagram where a single map contains all ensemble forecasts (e.g., Toth et al., 1997). However, many questions about how to best interpret and evaluate ensemble forecasts still remain.

This issue becomes even more complicated for mesoscale quantitative precipitation forecast (QPF), since QPF is a discontinuous field. Employing the technique of the ensemble mean and spread may yield meaningless results. More over, due to various errors, such as phase shift, rotation, and deformation errors (see Fig. 1), it has been a challenge to generate and interpret the ensemble QPF products.

Appropriate verification tools are essential in understanding the abilities and behaviors of an ensemble forecast system. Since each member of the ensemble is a deterministic forecast and therefore each member is subjected to systematic error (bias), it would be helpful in understanding the ensemble precipitation forecasting to examine and extract individual systematic error distributions. However, for conventional statistical and other methods it has been very difficult to properly take into account spatial characteristics of forecasting precipitation fields including phase shift, rotation, and deformation. This difficulty hinders further studies on spatial distributions and relationships.

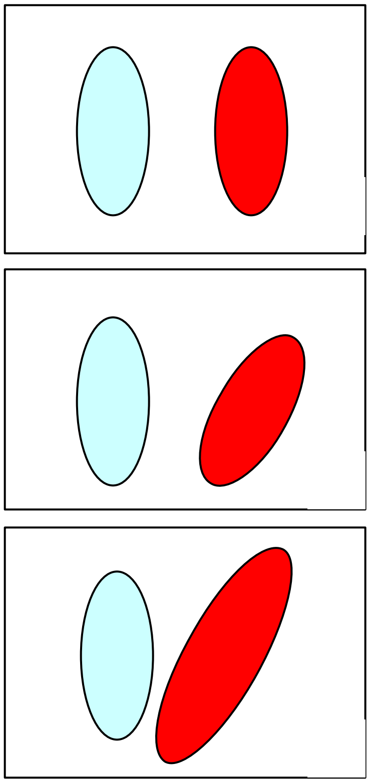

Fig.1 Illustration of three precipitation forecasting (deep gray shaded region) errors against observation (light gray shaded region): a) Forecast with phase shifting error; b) Forecast with shifting and rotation error; c) Forecast with shifting, rotation and deformation error.

Aimed at overcoming the above problems, this study proposes a spatial data mining approach as

[*] *Corresponding author address*: Xuechao Yu, Warning Decision Training Branch, 3200 Marshall Ave. Suite 202, Norman, Oklahoma 73072 email: Xuechao.Yu@noaa.gov

a new tool for verifying and understanding ensemble precipitation forecasts. Data mining is a process of exploring the data interrelationships or patterns from a potentially large volume of data using techniques that go beyond a simple search through the data (Huntsville, 1999; Fayyad, 1998). Data mining algorithms, in contrast to conventional methods, use discovery-based approaches in which pattern-matching and other algorithms are employed to determine the key relationships in the data. Data mining can look at numerous multidimensional data relationships concurrently, highlighting those that are dominant or exceptional. Since most of meteorological and geospatial data are related objects that occupy space, recent studies on data mining have extended the scope of data mining from relational and transactional databases to spatial databases (Koperski et al., 1996).

Spatial data mining can be classified into many categories based on data mining functionalities (Goebel and Gruenwald, 1999). For our specific study, we mainly focus on the following three aspects:

1) Pattern recognition

Although weather systems or features are well understood once shown graphically, many of them are ill-defined and hence difficult to quantify and extract using brute-force numerical methods. However, the identification and extraction of spatial objects share many similarities with the recognition and figure-ground separation problems in computer vision (Huang and Zhao, 1999). A spatial object is not just a collection of constituent objects but with a rich internal structure that may influence the aggregate properties of the object including its identity. Therefore, it is feasible to bridge the semantic gap between the input data of a massive grid field and the final symbolic description.

2) Spatial clustering

Clustering is usually one of the first steps in spatial data mining analysis and has been studied extensively for many years. The main advantage of using this technique is that interesting structures of clusters or natural grouping of distribution of data can be found directly from the data and help users to focus on a particular set of clusters for further analysis.

One of th3 well-known methods is the *Partitioning algorithm* which constructs a partition of a database D of n objects into a set of k clusters. k is an input parameter for these algorithms, i.e., some domain knowledge is required which unfortunately is not available for many applications. k-means algorithm is an example of *Partitioning algorithms*. The general weakness of partitioning-based algorithms is their inability to find arbitrarily shaped clusters.

It is often desirable to find natural clusters, i.e., clusters which are perceived as crowded together by human eye. To meet this need, other clustering methods have been developed based on the notion of *density* (Ester et al., 1996; Han et al, 2001).

The first density-based algorithm is the DBSCAN algorithm (Ester et al., 1996), which judges the density around the neighborhood of an object to be sufficiently dense if the number of data points within a distance ε of an object is greater than *MinPts* number of points. As the clusters discovered are dependent on the parameters ε and *MinPts*, DBSCAN relies on the user's ability to select a good set of parameters. To help overcome this problem, a cluster ordering method called OPTICS (Ankerst et al., 1999) was proposed. Rather than producing a data set clustering explicitly, OPTICS computes an augmented *cluster ordering* for automatic and interactive cluster analysis. One of the advantages of OPTICS is that OPTICS can handle very different local densities to reveal clusters in different regions. However, the required parameters of ε and *MinPts* still rely on user’s decision, since we may not see clusters of lower density.

In order to avoid any required parameters, in this research, we used Delaunay triangulation to get neighbors and local densities. We then adapted OPTICS cluster ordering idea to reveal clusters in different regions.

3) Spatial association rule

A spatial association rule is a rule which describes the implication of one feature or a set of features by another set of features in spatial databases. For example, a rule like “80% of schools are close to parks” is a spatial association rule.

To determine a specific spatial distribution region of an ensemble member, the concept of alpha shapes is applied, which is an approach to formalize the intuitive notion of "shape" for spatial point sets. The clustering points are used as the spatial point sets to determine their shape. Therefore, the spatial association rules between the observation and each ensemble forecast can be derived with a specific spatial distributed region.

In order to implement and test this approach, we generated a simulated dataset, in which observed and forecast precipitation regions are approximated by ellipses. Using the dataset, pattern characteristics or features, such as the center, axis and rotation, from the fields are depicted, and a new density-based clustering algorithm is then designed to extract the spatial precipitation distribution information. Finally, spatial data mining association rules are derived from the database to reveal the relationships among the observations and forecasts.

The organization of the rest of this paper is as follows. We will first introduce the simulated dataset in Section 2. Following this is an explanation of a pattern recognition method in Section 3. Sections 4 and 5 describe the proposed density-based clustering with Delaunay triangulation and spatial association rules. Discussion and conclusions are given in Section 6.

2. GENERATION OF SIMULATED DATASET

As discussed previously, for precipitation forecast, there are generally three kinds of errors involved: phase shifting, rotation and deformation. Therefore, in our study, generating a 2-D precipitation dataset that contains such errors is important and necessary. This error information will be used in clustering analysis and for further association rule analysis as well.

To simplify the problem of extracting patterns, with the possibility of generalization later, we use ellipses to represent the patterns in the precipitation field. The standard ellipse equation is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad (1)$$

where *a* and *b* *are* the half axes. Rotation of a point (*x*, *y*) with respect to the origin is given by

$$x' = x * \cos \varphi - y * \sin \varphi, \qquad (2)$$

$$y' = x * \sin \varphi + y * \cos \varphi, \qquad (3)$$

where φ is the angle of rotation. Furthermore, we need to consider the shift (x_c, y_c) from the origin. For a given center of ellipse (x_c, y_c), the horizontal and vertical axes *a* and *b*, and a rotation angle (φ), we can specify an ellipse. For our test, we use a 200 × 200 domain and set precipitation to zero outside the ellipses and to 1" inside. We also assume that each model forecast only contains one precipitation field. Figure 2 illustrates the samples of our datasets. The dataset contains forecasts from 20 member models and each model generates 100 forecasts.

3. PATTERN CHARACTERISTICS SELECTION

The spatial objects or features perceived by scientists have common characteristics: they are all visually salient. Therefore, the symbolic description must impose a conceptual structure on the system so that the complexity of the system can be understood in terms of well-defined parts and subparts, and the interactions among them. It is important to represent properly and objectively some specific properties of the objects with aggregate spatial properties. We choose the center, major and minor axis and its rotation for the symbolic descriptions. Moreover, additional specific properties for further descriptions with a given classified object are developed. For example, for a closed space curve, except for characteristics like center, major and minor axis, additional properties and constrains are needed to distinguish whether the object is similar to an ellipse, circle or square. With this feature selection technique, it becomes feasible to measure the errors of phase shifting, rotation and deformation for further data mining.

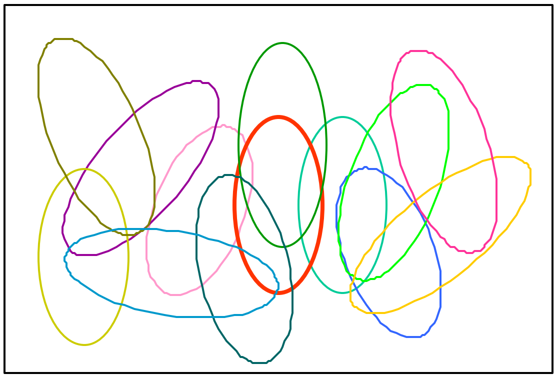

Fig.2 Multiple model precipitation forecasts and observation (dark and thick-line ellipse)

This mining process will result in the spatial data in terms of the center (*x*, *y*), minor and major axis (*a*, *b*) and rotation angle. Adding the model number and the sample number information, the output of relative forecasting information of phase shift, azimuth and rotation against observation will be like:

(modelNo, sampleNo, x_center, y_center, minor_axis, major_axis, rotation).

4. DENSITY-BASED CLUSTERING WITH DELAUNAY TRIANGULATION

The Delaunay triangulation of a point set is a collection of edges, which satisfies an "empty circle" property: for each edge we can find a circle containing the edge's endpoints but not containing any other points (Cetin, 2000). With this characteristic, Delaunay triangulation becomes an excellent method for solving the nearest neighbor problem, because each point in the diagram is connected to its nearest neighbor (Fig. 3). For a point set, p_1, p_2, …, p_n, let $NDE(p_i)$ denote the number of Delaunay edges incident to p_i. Therefore, the local mean distance or edge of a point and global mean distance can be calculated by the following equations:

$$LocalMean(p_i) = \frac{1}{NDE(p_i)} \sum_{j}^{NDE(p_i)} d_i \qquad (4)$$

$$GlobalMean = \frac{1}{n} \sum_{i}^{n} LocalMean(p_i) \qquad (5)$$

where d_j is the length of Delaunay edge. We use the *GlobalMean* as a radius to get the local density, *density*(p_i), which is the number of points within a circle centered at point p_i. To create an ordering of the object, instead of using *core-distance* and *reachability-distance* as in OPTICS (Ankerst et al., 1999), we search the point with minimum distance, d_i, among the points not yet searched and record the local density, *density*(p_i) (Fig. 4).

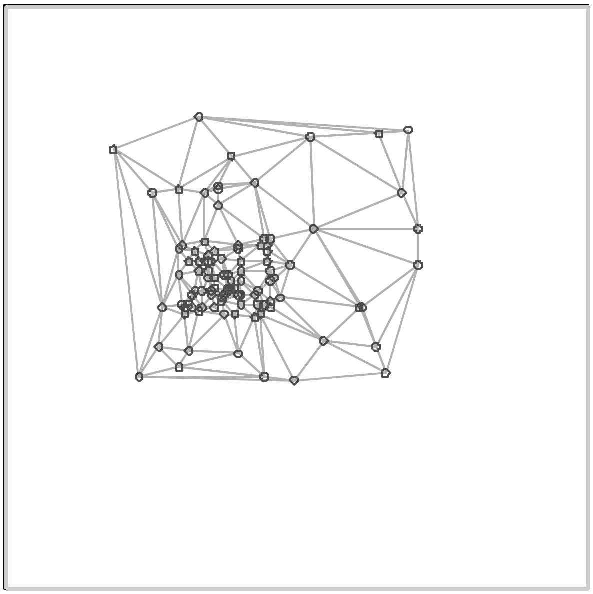

Fig.3 Illustration for Delaunay triangulation of 100 samples of model No. 4.

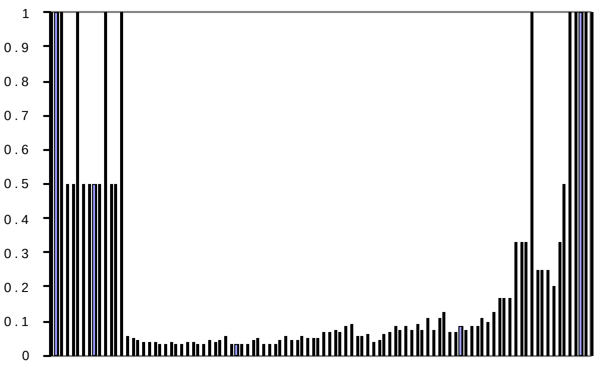

Fig.4 Cluster ordering of Model 4.

With the ordered points, one can obtain various clustering results, which are constrained by different thresholds. For example, in this study, we use maximum polygon area, *Maxs*, and minimum confidence, *Min c%*, to be introduced in the next section, as the specific constraints in our clustering analysis. Those points not satisfying the thresholds of constraints are viewed as noises. Figure 5 shows one of the clustering results, in which the solid gray circles are clustering points and others are noises.

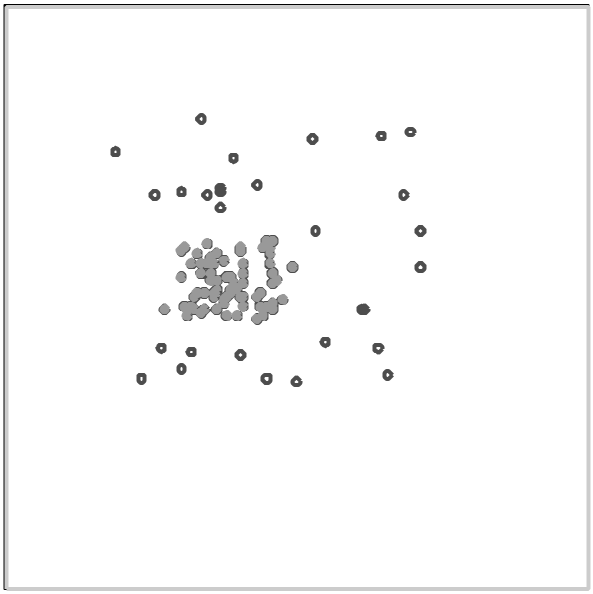

Fig.5 Clustering results of model No. 4. Solid gray circles are those points satisfying the threshold and constraints.

5. SPATIAL ASSOCIATION RULE ANALYSIS

In order to discover the knowledge about the hidden relationships among the observations and forecasts, spatial association rules, which can be derived through the data mining and provide very important information. An association rule is a rule that associates one or more spatial objects with other spatial objects, which is of the form X→ Y (c%), where X and Y are sets of spatial or non-spatial predicates and c% is the confidence of the rule.

There are various kinds of spatial predicates that could constitute a spatial association rule. Examples include: spatial orientations like *left_of*, *west_of*, *etc.*; and distance information, such as *close_to*, *far_away*, etc.

With the spatial distribution of individual model forecasts revealed by our density-based clustering algorithm, it is possible to quantitatively describe the location, size and shape of the clustering objects for spatial association rule analysis. The concept of alpha shapes formalizes our intuitive notion of "shape" for a spatial point set of data. Alpha shapes can be viewed as generalizations of the convex hull of the point set (Edelsbrunner and Mucke, 1994). It formalizes the intuitive notion of shape, and for varying parameter alpha, it ranges from crude to fine shapes. The most crude shape is the convex hull itself, which is obtained for very large values of alpha. As alpha decreases, the shape shrinks and develops cavities that may join to form tunnels and voids. Fig. 6 represents an alpha shape for the clustering data. The alpha shape forms a polygon region, which indicates that one ensemble member forecasts appearing in such area with respect to the observation have a higher probability.

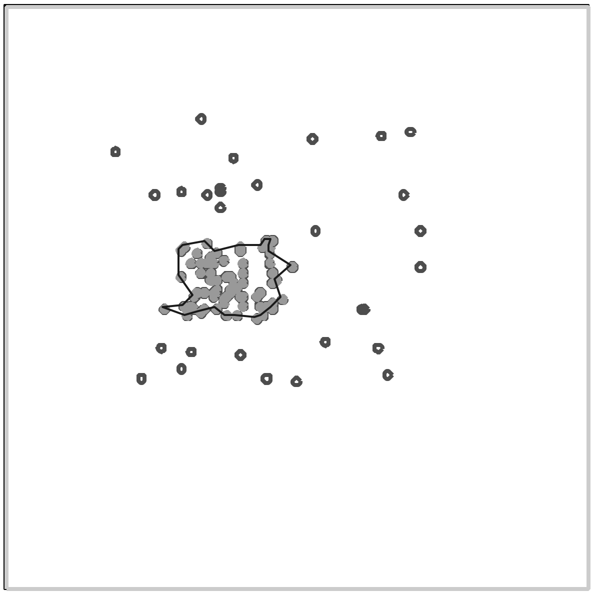

Fig.6 Alpha shape for clustering results of Model 4. Solid gray circles are those points satisfied the threshold and constrains.

To confine the number of discovered rules, the concept of minimum confidence is used. The confidence is the proportion of database transactions containing X that also contain Y. The intuition behind this is that in large databases, there may exist a large number of associations between objects but most of them will be applicable to only a small number of objects, or the confidence of rules may be low. We are only interested in strong rules, which are rules with large confidence, i.e., no less than the minimum confidence threshold, *Minc%*. With the alpha shaped polygon, we first calculate the polygon area, s, and confidence, c%. Then we check if the *c%* > *Minc%*, and s < *Maxs*. In the study, *Minc%* is 60% of all samples. This threshold should be determined and verified based on real case studies.

Therefore, spatial association rules can be derived and represented as "model(*X*) → event_in(polygon P_x) (c%)". Since the polygon location is a relative location with respect to the observation, we can further interpret it with information such as, distance (close_to, far_away_to,…), azimuth (northeast_of, west_of, …), and orientation (normal, left_bias, …), etc. This information reveals each ensemble member's spatial distribution and behavior. These spatial association rules will help us in evaluating and understanding the ensemble precipitation forecasts.

6. DISCUSSIONS AND CONCLUSIONS

This paper proposed a spatial data mining approach as a new tool for verifying and understanding ensemble precipitation forecasts. With this approach, a particular attention is given to the spatial distribution characteristics of precipitation. Two particular issues are addressed in this paper: first the assessment of forecast quality through pattern recognition which can identify errors due to phase shift or displacement, rotation and deformation, and second the investigation of association among observations and forecasts.

With the feature selection process, spatial characteristics such as center of forecasting precipitation field, axis, and rotation can be selected. These results make it possible and convenient to depict quantitatively the spatial distribution of each ensemble member. Our proposed density-based clustering algorithm using Delaunay triangulation can automatically find out natural clusters of ensemble members' precipitation forecasts. Moreover, the proposed algorithm allows users to apply additional constraints, such as confidence and polygon area, for interested clusters. The alpha shape concept is applied to represent graphically the shape of the clusters.

Based on the knowledge gained from the distributions, a further step is taken to derive association rules between the observations and forecasts from the database. Eventually, the approach provides information on the distribution of individual ensemble members and reveals the relative spatial associations among the observations and forecasts with a certain confidence. In our experience with our simulated dataset of 20 ensemble members , most members have a strong association rule with observations, although their spatial distributions vary in a wide range. Some ensemble member forecasts do not have a strong association rule with the observation, which implies that these members have less value in the ensemble forecast.

Therefore, this approach can not only takes into account the spatial distribution characteristics of individual ensemble members, which are otherwise difficult to describe, especially quantitatively, with conventional composite charts or "spaghetti" charts, but also provides strong spatial association rules between the observation and each ensemble member. These properties can help users in verifying and understanding the ensemble precipitation forecasting. Since the spatial association rules indicate systematic errors for each member and provide confidence in specific polygon region, users can make use of the information in generating and interpreting ensemble products.

It should be mentioned that we used a simulated dataset so that we can focus on the development of the procedure for mining and deriving spatial characteristics, distributions and relationships of individual ensemble members. Extending this procedure to realistic precipitation data sets is our next step. We envision that by adding more shapes and characteristic representations, we will be able to apply the procedure to real ensemble forecasts to reveal much valuable information so as to help us understand , verify and interpret the ensemble forecast.

7. REFERENCE

Ankerst, M., M. Breunig, H. -P. Kriegel, J. Sander: OPTICS: Ordering Points To Identify the Clustering Structure, Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA, 1999, pp. 49-60.

Cetin, N., 2000: A Local Optimization Algorithm for Well-Graded meshes. Thesis, http://www.inf.ethz.ch/personal/cetin/thesis/thesis/node2.html.

Edelsbrunner, Herbert and E. P. Mucke. 1994. Three-dimensional alpha shapes, ACM Trans. Graphics, 13, 43-72.

Ester, M., H. -P. Kriegel, J. Sander, and X. Xu, 1996: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, 1996, pp. 226-231.

Fayyad, Usama, 1998: Mining database: Towards algorithm for knowledge discovery. *Data Engineering*. Vol. 21 No. 1, March 1998.

Goebel, M. and L. Gruenwald, 1999: A survey of data mining and knowledge discovery software tools. *SIKDD Explorations*, Vol. 1, Issue 1, June 1999, pp. 22-33.

Han, J., M. Kamber, and A. K. H. Tung 2001, Spatial Clustering Methods in Data Mining: A Survey, H. Miller and J. Han (eds.), {sf Geographic Data Mining and Knowledge Discovery}, Taylor and Francis, 2001.

Hou, Dingchen, E. Kalnay, and K. K. Droegemeier, 2001: Objective Verification of the SAMEX '98 Ensemble Forecasts. *Mon. Wea. Rev*., 129, 73–91.

Huang, X. and F. Zhao, 1999: Seeing Objects in Spatial Datasets. Proc. of 3rd Int'l Symp. on Intelligent Data Analysis (IDA-99), Amsterdam, August, 1999.

GSFC, 1999: NASA workshop on issues in the application of data mining to scientific data. *Final Report*. Goddard Space Flight Center, Greenbelt, Maryland.

Koperski, Krzysztof, J. Han and J. Adhikary, 1996: Spatial data mining: Progress and challenges. SIGMOD'96 Workshop. on Research Issues on Data Mining and Knowledge Discovery (DMKD'96), Montreal, Canada, June 1996.

Toth, Z., and E. Kalnay, S. Tracton, R. Wobus, and J. Irwin, 1997: A synopticevaluation of the NCEP ensemble. *Wea. Forecasting*, 12, 140–153.