A Distance Function Based Algorithm to Quantify Uncertainty in Areal Limits

Amir H Hosseini and Clayton V. Deutsch

Department of Civil and Environmental Engineering
University of Alberta

Uncertain 'Areal limits' are problematic in many geo-engineering applications. Geostatistical techniques algorithms have been developed to solve various spatial modeling problems. These techniques deal with variations of discrete and continuous variables 'within' the domain of interest. They are not particularly efficient when it comes to quantifying uncertain areal limits. In this paper, a Distance Function (DF) based algorithm is introduced to provide a framework for delineation of uncertainty for 2-dimensional areal limits. The distance function approach considers measurements of the presence or absence of a particular attribute across the domain. In the proposed methodology, a large number of synthetic limits are generated and then used to calibrate a 'band of uncertainty' for areal limits. A cross-validation exercise is then implemented to assess the performance of the methodology. The 'band of uncertainty' resulted from this procedure can be used to draw equi-probable realizations of areal extents. These realizations can be easily used in a Monte Carlo Simulation framework to bound realizations of properties within the site.

Introduction

In environmental applications, a threshold is often defined to create a sharp boundary between contaminated and uncontaminated locations. Under steady-state condition and for a pre-specified threshold, there are some areas that are clearly contaminated and some areas are clearly clean. In other words: a sharp boundary exists between contaminated and uncontaminated areas.

Geostatistical approaches (Deutsch and Journel 1998) and Machine-Learning algorithms (Kanevsky et al. 1996) have been developed and frequently used to solve various stationary and non-stationary problems. Non-stationarity in data is usually handled by decomposing the data into a trend and a residual:

$$Z(u) = m(u) + R(u) \tag{1}$$

The trend component is often modeled deterministically. Fitting polynomials, inverse distance (ID) interpolation, block kriging (Deutsch 2002) and Neural Networks (Kanevsky et al. 2004) are some of trend modeling approaches used in practice. The residual component is modeled by geostatistical techniques. The above approaches have shown lots of promise in modeling earth science attributes at unsampled locations. However, they do not provide a reasonable measure of uncertainty particularly at the edges, when a sharp boundary is expected.

In order to deal with this problem, Froidevaux et al. (2001) used an analytical model of transport to generate a series of contaminant concentration maps, each map resulted from a different yet equiprobable set of parameters. This suit of concentration images is then summarized into an a-priori contamination probability map, describing the probability that, at any location, the contamination concentration exceeds a critical threshold. This prior probability is then conditioned to available hard data (piezometer readings), using indicator simulation with locally varying means. One drawback of this methodology is that it can not be used in complex situations such as characterization of sites impacted with residual NAPL. This is due to the fact that the distribution of residual NAPL in soil can not be explained by any analytical or numerical model, as it is a complex function of the soil texture and groundwater fluctuations. There is another drawback for this methodology: too much weight is assigned to the a-priori map obtained by Monte Carlo Simulations, which has been considered as a locally varying mean for the subsequent indicator kriging.

Gleyze et al. (2001) also proposed an extension to Wombling procedure (Womble 1951) to characterize the regions of abrupt change of a spatial variable *Z* for which a finite number of data are available on a given geographical area. Although this method has shown some promise (especially in biological sciences community), it requires a statistical test and assumptions about significance of the detected barriers that make it ‘modeler-dependent’ for a large part.

The distance function based approach proposed in this paper is an attempt to overcome the above drawbacks and develop a simple and robust methodology for characterization of areal extents in a probabilistic framework.

The space of uncertainty for areal limits can be represented by an ‘uncertainty band’, as shown in figure (1 – a, b). For every data location, Distance Function (DF) parameter is defined as the distance to the nearest ‘unlike’ data location. It is positive, when a data location meets a pre-specified condition (e.g. contaminant is present) and negative otherwise. The centerline and bandwidth of the uncertainty band should be determined by modifying the Distance Function parameters. The DF parameters should be systematically modified until the centerline and width of the uncertainty band are calibrated against a large number of synthetic plumes.

The Distance-Function based algorithm is very simple in concept and consists of a few steps. Its implementation details come in the next paragraphs. It can be used to characterize the uncertainty in the areal limits of any earth-science attribute. The characterized space of uncertainty for areal limits can be easily used in a Monte Carlo Simulation framework to clip realizations previously obtained by geostatistical analysis (Hosseini et al. 2006). In order to evaluate the methodology, a cross-validation exercise is also implemented.

Step 1: Assigning the control points and calculating the DF values

First, a number of user-defined control points are assigned at locations that are certainly uncontaminated. Figure 2 – a shows the configuration of contaminated/uncontaminated wells and figure 2 – b shows the locations of control points and the distance function values calculated for all data locations and control points.

Step 2: Mapping the distance function values

Inverse distance interpolation can be used to map DF values across the domain. An inverse distance (ID) interpolation is a spatially weighted average of the sample values within a search radius. It is calculated as:

$$Z^*(u) = \sum_{i=1}^{N} \lambda_i Z(u_i) \tag{2}$$

in which, u is the unsampled location, $Z^*(u)$ is the ID interpolation estimate at the unsampled location, λ_i ’s are weights assigned to each conditioning data at sample points, and N is the number of data locations in the search neighborhood. The weights can be calculated by:

$$\lambda_i = \frac{\left(\frac{1}{d_i^{\omega}}\right)}{\sum_{i=1}^{n}\left(\frac{1}{d_i^{\omega}}\right)}, \quad (i = 1, \ldots, n) \tag{3}$$

in which, d_i ’s are Euclidian distances between estimation location and sample points. Exponent ω is the distance exponent value and can take any value from 0.5 to 2. ω is traditionally calibrated by cross validation. Recently, Muller et al. (2005) showed that cross-validation may not give the most appropriate value for the distance exponent value, as the distances between the prediction points are usually much larger than those of sampling locations. They suggest that a more reliable approach would be to use a distance exponent value between 1.5 and 2.0. Nevertheless, ID interpolation should still be optimized by limiting the number of data (the search neighborhood) used in the interpolation (Rojas-Avellaneda and Silvan-Cardenas 2006). As a result, the ID interpolation weights can be calculated by:

$$\lambda_i = \begin{cases} \dfrac{\left(\dfrac{1}{d_i^{\omega}}\right)}{\sum_{i=1}^{n}\left(\dfrac{1}{d_i^{\omega}}\right)} & d_i \le r \\ \\ 0 & d_i \le r \end{cases} \tag{4}$$

where, *r* is the radius of the search neighborhood, which should be calibrated by cross validation.

Figure 3 shows the contour line of DF = 0 determined by ID interpolation of DF's with a distance exponent value ω of 1.5 and a calibrated search radius. This contour line represents an initial guess for the plume boundary.

Step 3: Generating multiple synthetic plumes

Multiple synthetic plumes may be generated by locating random wells across the modeling domain. First, a search angle α is specified by the modeler. Then, for every contaminated well a search is implemented and all the search directions that include a closest 'unlike' data location are identified as 'valid' search directions and those include a closest 'like' data location are identified as 'null' search directions. In order to generate a synthetic plume, a search direction is randomly selected for each contaminated well. The selected search direction can be either a 'null' or a 'valid' search direction. If a 'valid' search direction is selected, a new 'imaginary' well is added to the domain. This 'imaginary' well is randomly located on a line that connects the original contaminated well location to the closest 'unlike' well location. The new imaginary well is randomly assigned to be either contaminated or uncontaminated. Repeating this process generates multiple synthetic plumes that can be used in calibration of uncertainty band. Figure 4 schematically shows the process. The search angle α is very important in this process. In fact as α decreases, more short scale variations appear in the generated synthetic plumes. Figure 5 shows a number of synthetic plumes with a search angle α of 30°.

Step 4: Calibrating the band of uncertainty

The centerline and width of uncertainty band are calibrated by (1) systematically modifying the DF values for all data locations and control points, (2) mapping the new set of DF values across the domain and (3) minimizing an objective function to find the best set of DF values. Modifying the DF values has been made possible by introducing parameters C_1 and C_2. For every attempt, original DF values are modified by adding or subtracting a new C_2 value and multiplying to or dividing by a new C_1 value:

$$DF_{new} = \begin{cases} \left(DF_{old} + C_2\right) \times C_1 & DF_{old} \ge 0 \\ \\ \left(DF_{old} - C_2\right) / C_1 & DF_{old} < 0 \end{cases} \tag{5}$$

Then, the new set of DF values is mapped across the domain by ID interpolation and a new uncertainty band is obtained. Changing C_1 and C_2 values will result in changing in the location of the centerline and width of uncertainty band. Figure 6 shows the uncertainty band for different C_1 and C_2 values. To calibrate DF values at different data locations, one needs to find C_1 and C_2 values to minimize the objective function:

$$OF = \sum_{j=q_1}^{q_M} \sum_{i=\alpha_1}^{\alpha_N} \left(P_j^{True} - P_{i,j}^{Calc}\right)^2 \tag{6}$$

in which, P_j^{True} are the probabilities corresponding to quantiles q_1 to q_M that are used in calibration, $P_{i,j}^{Calc}$ is calculated as the proportion of synthetic plumes that 99 % of their area is covered by the quantile map corresponding to a given uncertainty band and a given quantile from the set of q_1 to q_M.. The uncertainty band is calculated each time with a different set of C_1 and C_2 values. α_1 to α_M are the search angles that have been used to generate the synthetic plumes. A reasonable number of quantiles and a reasonable number of search angles must be used in calibration. The quantiles q_{10}, q_{30}, q_{50}, q_{70}, and q_{90} and the search angles of 5°, 10°, 30°, 45°, 60°, and 90° have been used in this study.

Figure 7 represents the *OF* for different C_1 and C_2 values; and depicts the C_1 and C_2 values that minimize the objective function are: $C_1 = 1.36$ and $C_2 = 12.92$. The optimized uncertainty band and the corresponding p_{10} to q_{90} maps for the calibrated C_1 and C_2 values are shown in figure 8.

Performance assessment through cross-validation

In order to assess the performance of the proposed methodology, a cross-validation exercise is implemented. Some of contaminated and uncontaminated wells were removed from the system one by one (with replacement), and the probability of being contaminated was calculated for each well based on the proposed methodology. Results are shown in table 1.

Table 1: Cross-validation results for 11 wells deemed closely located on the boundaries of the plume

Well ID	C1	C2	Target Probability	Calc. Probability
1	1.327	14.11	1	1
2	1.331	12.94	1	1
4	1.31	13.8	1	0.8339
7	1.28	18.3	1	0
8	1.375	14.1	1	0.2216
9	1.345	13.82	1	1
13	2.257	41.31	0	0.01291
14	2.481	43.43	0	1
15	2.491	38.91	0	0
16	2.356	46.88	0	0
18	2.321	45.55	0	0.8852

As it can be observed in table 1, cross-validation shows that the proposed methodology can reasonably detect the areas with potential contamination and exclude the areas that are potentially clean. The methodology, however, fails to detect the contaminated areas when some discrepancies are present, meaning one or more contaminated wells exist far away from the main body of the plume. This is the case for well 7, in the above example.

Conclusions

Delineation of uncertainty in areal limits is very important and challenging in many science and engineering applications. Defining the ‘Distance Function’ concept, a simple and powerful methodology was proposed to characterize the space of uncertainty for areal limits. The methodology was tested by cross-validation and deemed to be robust, unless some discrepancies exist such as a contaminated well far away from the main body of the plume. The model of uncertainty obtained by this methodology can be easily incorporated in Monte Carlo Simulation to simulate any continuous attribute with uncertain areal limits.

References

Deutsch, C.V. and Journel, A.G.: GSLIB: Geostatistical Software Library and Users Guide, Oxford University Press, New York, second edition, 1998.

Deutsch, C.V.: Geostatistical Reservoir Modeling, Oxford University Press, New York, 2002.

Froidevaux, R., Garcia, M. and Blondel, T.: A probabilistic and conditional approach for delineating the extent of a contamination plume within an aquifer. in Monestiez P., Allard D. and Froidevaux R. Eds, geoEnV III – Geostatistics for Environmental Applications, Kluwer, 217 – 226, 2001.

Gleyze, J.F., Bacro, J.N. and Allard, D.: Detecting regions of abrupt change: Wombeling procedure and statistical significance. in Monestiez P., Allard D. and Froidevaux R. Eds, geoEnV III – Geostatistics for Environmental Applications, Kluwer, 311 – 322, 2001.

Hosseini, A.H., Biggar, K.W., Deutsch, C.V. and Mendoza, C.A. Geostatistical Analysis of CPT-UVIF data for development of a Site Conceptual Model. Proceedings of NGWA conference on Petroleum Hydrocarbons and Organic Chemicals in Groundwater; November 6 – 7, 2006, Houston, TX, USA.

Kanevsky, M., Arutyunyan, R., Bolshov, L., Demyanov, V. and Maignan, M.: Artificial neural networks and spatial estimations of Chernobyl fallout, Geoinformatics, 7, 5 – 11, 1996.

Kanevsky, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V. and Canu, S.: Environmental data mining and modeling based on machine learning algorithms and geostatistics, Environmental Modeling and Software, 19(9), 845 – 855, 2004.

Mueller, T.G., Dhanikonda, S.R.K., Pusuluri, N.B., Karathanasis, A.D., Mathias, K.K., Mijatovic, B. and Sears, B.G.: Optimizing Inverse Distance Weighted Interpolation with Cross-Validation, Soil Science, 170(7), 504 – 515, 2005.

Rojas-Avellaneda, D. and Silvan-Cardenas, J.L.: Performance of geostatistical interpolation methods for modeling sample data with non-stationary mean, Stochastic Environmental Research and Risk Assessment, 20(6), 2006.

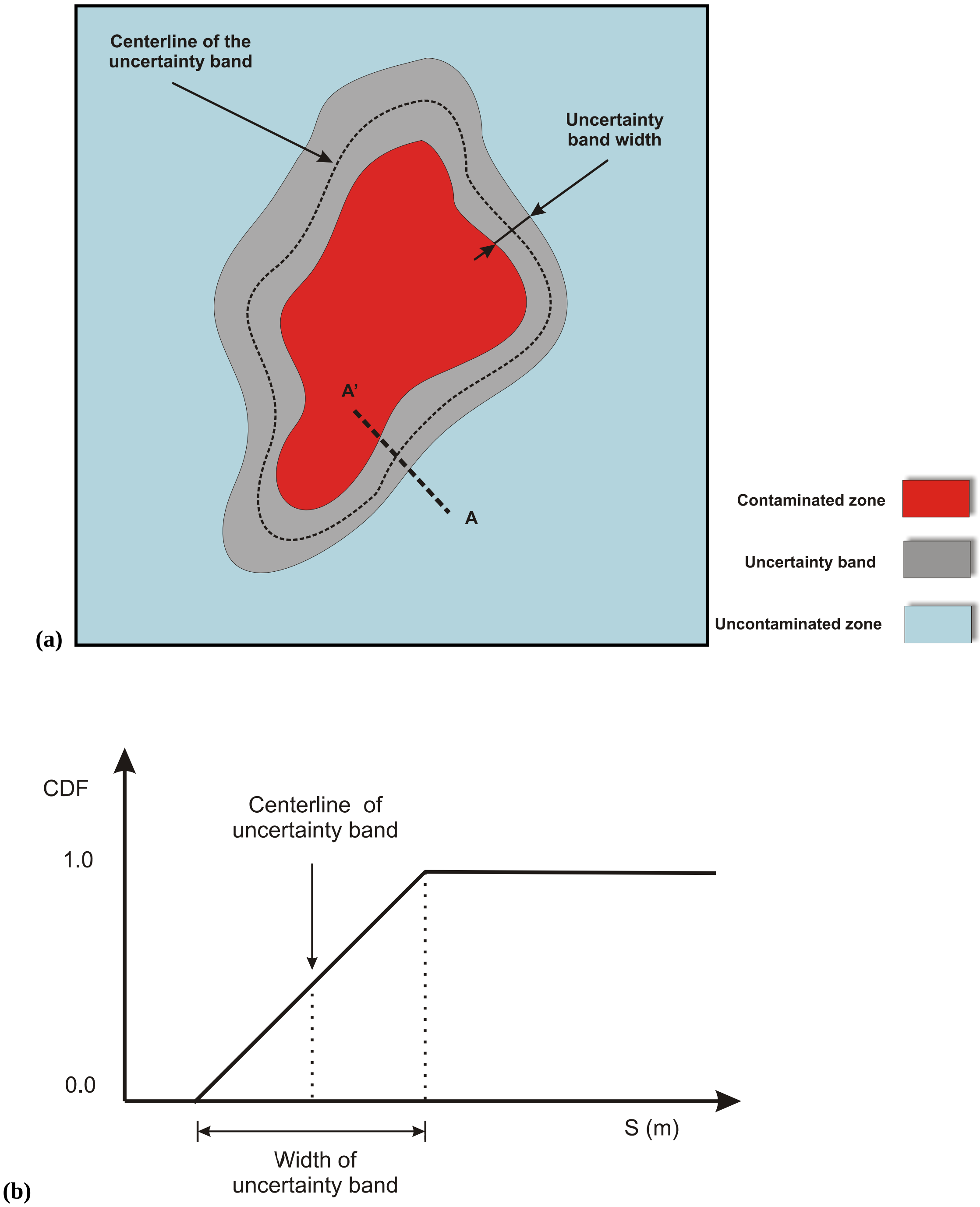
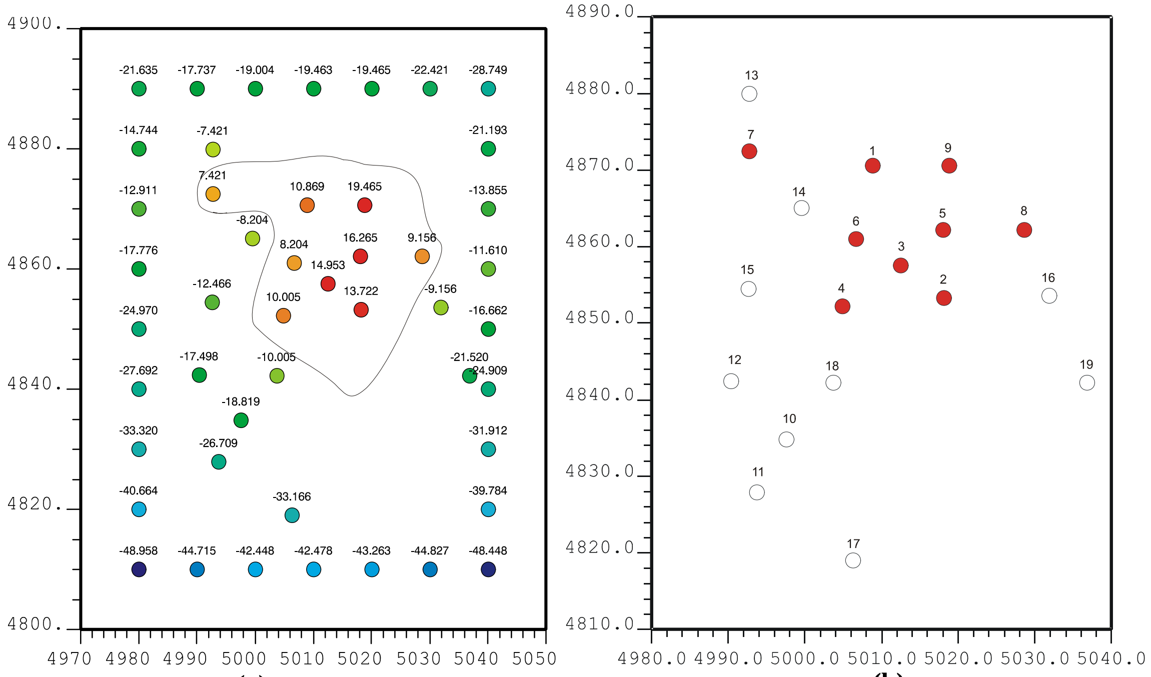


Figure 1: (a) A schematic uncertainty band and the corresponding uncertainty band width and centerline; (b) Cumulative probability distribution along A-A' cross-section



(a) **(b)**

Figure 2: (a) Configuration of contaminated (solid circles) and uncontaminated wells; (b) Designation of control points and calculation of DF values at all data locations and control points.

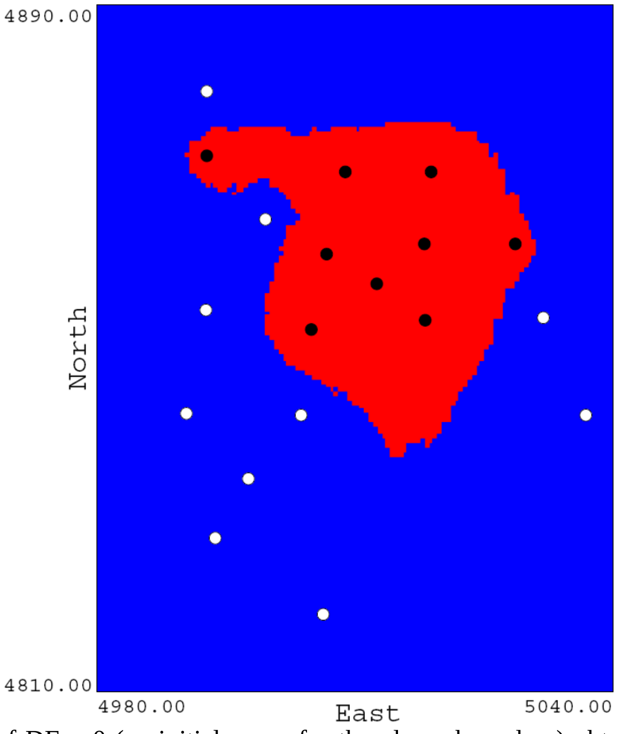


Figure 3: Contour line of DF = 0 (an initial guess for the plume boundary) obtained by ID interpolation with a distance exponent value ω of 1.5 and an optimized search radius.

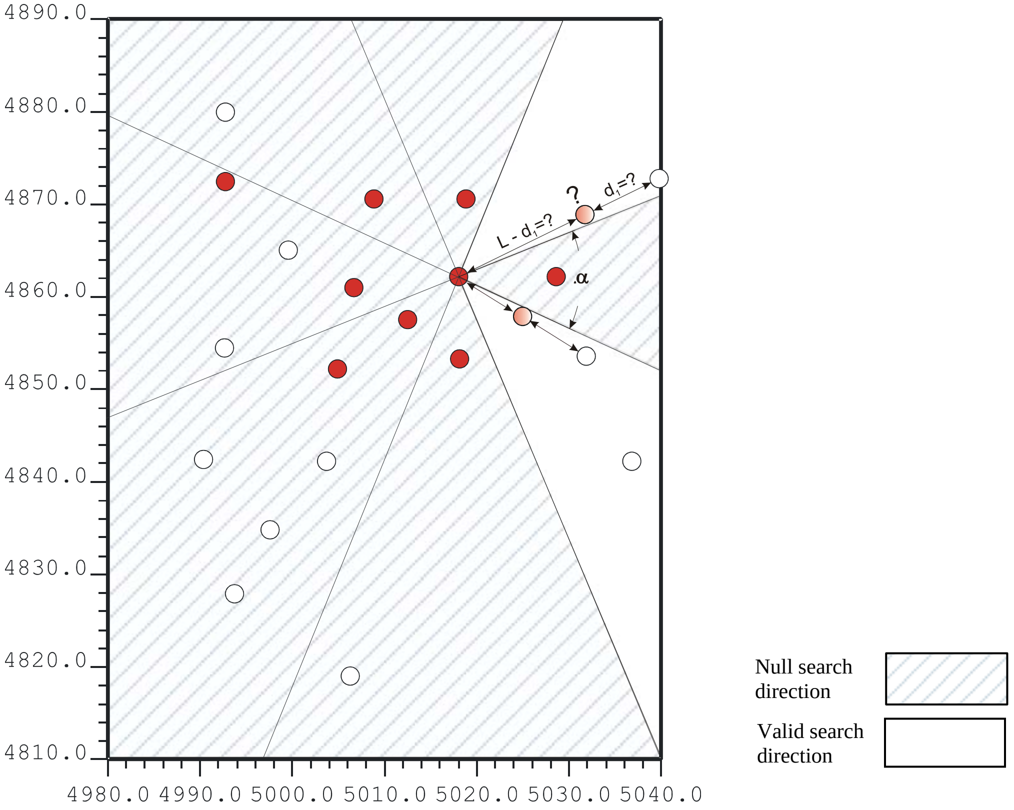


Figure 4: Procedure of generating multiple synthetic plumes to be used in calibration of uncertainty band.

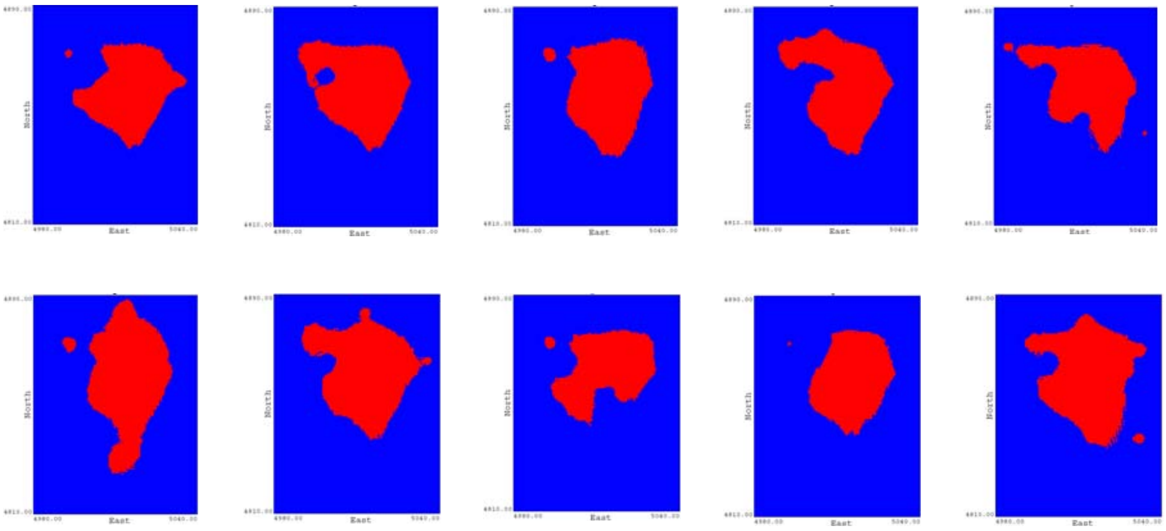


Figure 5: Synthetic plumes generated using the explained procedure with a search angle of $\alpha = 30°$.

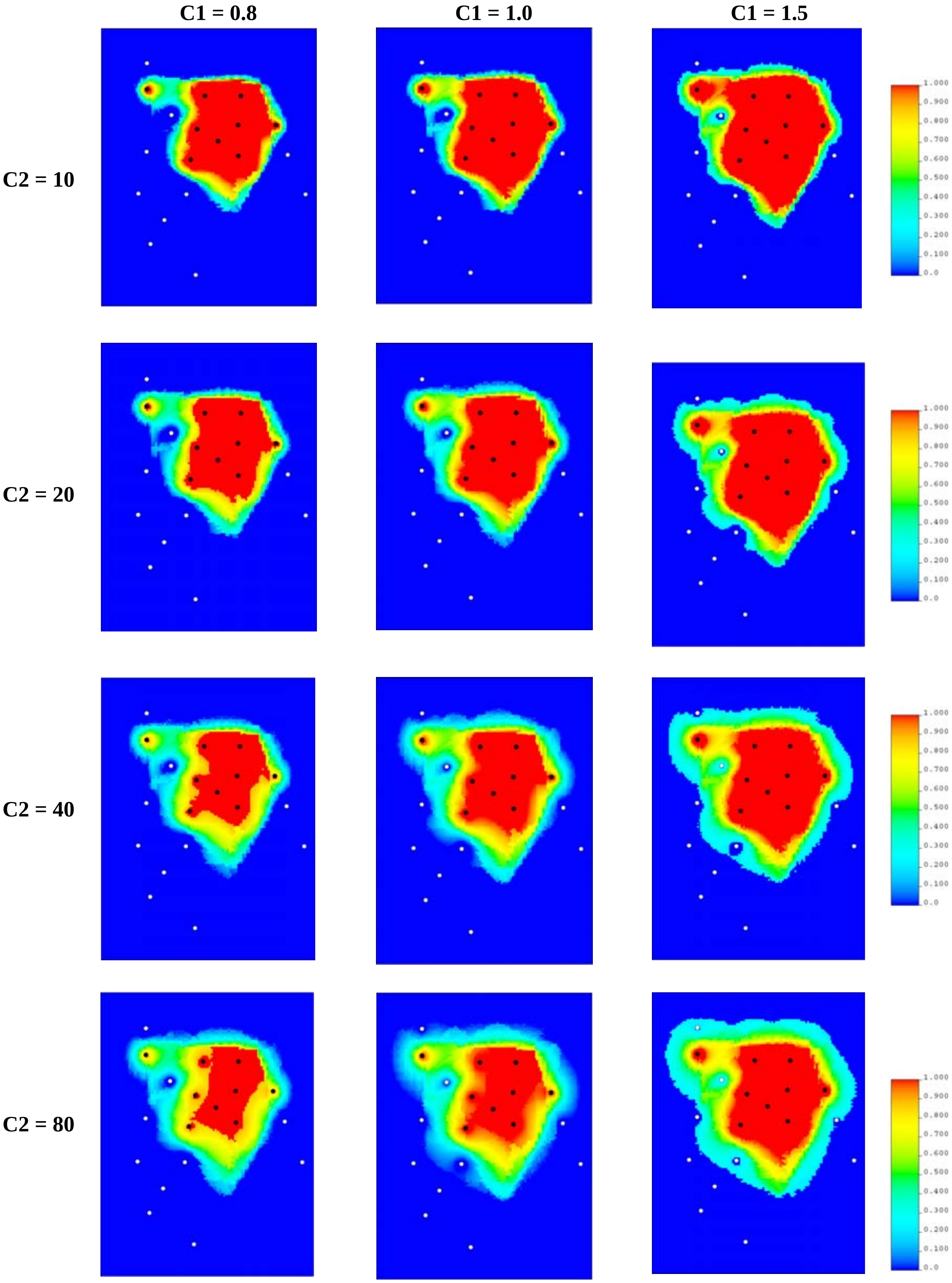


Figure 6: Different uncertainty bands as a function of C_1 and C_2 values.

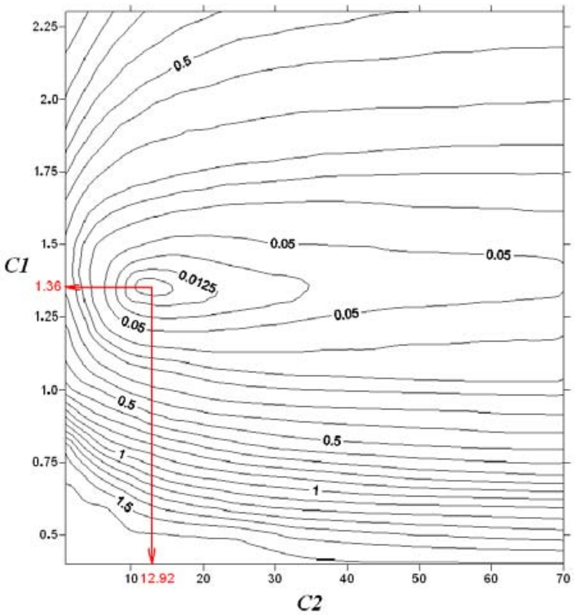


Figure 7: Objective Function as a function of C_1 and C_2 values. The C_1 and C_2 values that minimize the objective function are also shown.

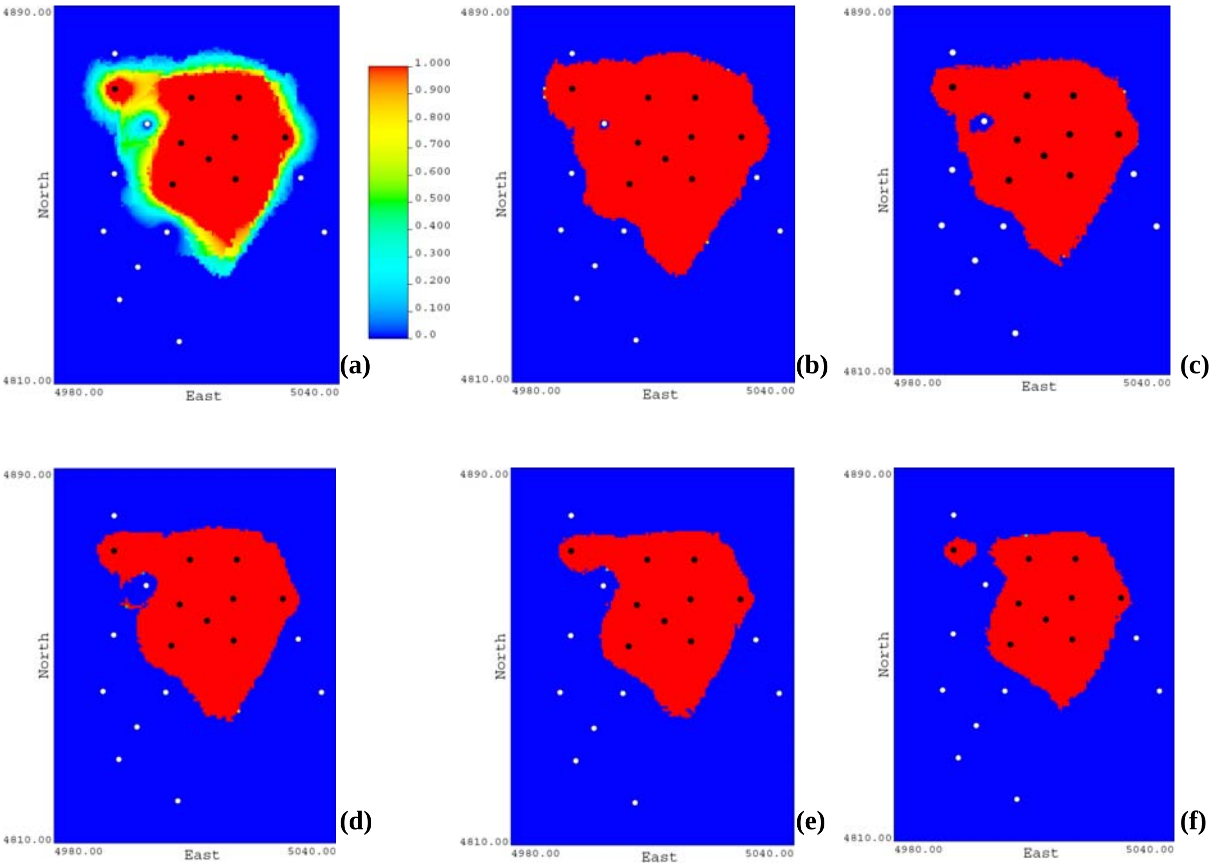


Figure 8: The optimized uncertainty band (a) and corresponding p_{10}, p_{30}, p_{50}, p_{70}, and p_{90} maps (b-f) based on the calibrated C_1 and C_2 values: 1.36 and 12.92, respectively.