A New Way to Calibrate Distance Function Uncertainty

Brandon J. Wilde and Clayton V. Deutsch

Perhaps the most critical decision in geostatistical modeling is that of choosing the stationary domains or populations for common analysis. The boundaries between the stationary domains must be modeled with uncertainty. Interpolating a distance function is a useful method for modeling boundaries with uncertainty. The current implementation of distance function boundary modeling with uncertainty requires expensive calibration with simulated data and numerous reference models to ensure unbiasedness and fair uncertainty. A method for using the available data to calibrate the distance function is proposed which greatly reduces the calibration expense.

Introduction

Interpolating a distance function has been shown to be a reasonable method of locating boundaries as it is simple and flexible (McLennan, 2008; Hosseini, 2009). However, it needs a large amount of hard data and does not provide direct access to uncertainty. The definition of the distance function is related to the notion of distance to an interface separating two distinct domains. Distance is measured to the nearest unlike data location. Distance can be positive or negative depending on the location of the data inside or outside the domain. Thus, a first guess for the bounding interface of interest would be the line or surface corresponding to a constant value of zero. The distance function varies smoothly between increasingly positive values outside and further away from the boundary interface to increasingly negative values inside and further away from the boundary surface. To determine the location of the boundary, the distance to the nearest unlike sample is calculated for all available samples. This distance function data is then used to condition the interpolation of distance function on a regular grid. The boundary is considered to lay at the transition between positive and negative interpolated distance function values.

There is uncertainty in the boundary location. Munroe and Deutsch (2008 a,b) propose assessment of the uncertainty by calibration of parameters, C and β where C controls the width of the uncertainty and β controls the bias. These parameters are optimized to give appropriate uncertainty. Optimizing these parameters is an expensive operation requiring multiple reference models and two objective functions.

This work proposes a simpler, less expensive calibration. Only the C parameter is calibrated and this is done using only the data to calibrate in a relatively computationally inexpensive manner. A subset of data is removed prior to the calculation of the distance function at the data locations. The subset of data removed will hereafter be referred to as the jackknife data. This effectively creates two data sets: the distance function data and the jackknife data. The distance function data are used to condition the interpolation of the distance function. The jackknife data are then considered. A number of jackknife data that are coded as inside the domain will have positive distance function estimates (outside the domain) and a number of jackknife data that are coded as outside the domain will have negative distance function estimates (inside the domain). The C parameter is adjusted until the desired proportion of incorrectly classified jackknife data are correctly classified. Once the C parameter is determined, it is applied to the calculation of the distance function for all available sample data. All of the data are then used to condition the interpolation of the distance function.

Consider the top plot in **Figure 1** where the distance function subset of the data has been used to determine the boundary represented by the black line where the distance function equals zero, the green dots represent data coded outside the domain, and the red dots represent data coded inside the domain. Consider the middle plot in **Figure 1** where the jackknife data have been reintroduced. The jackknife data coded outside are purple; the jackknife data coded inside are orange. There are a number of jackknife data coded outside that fall inside the domain and a number of jackknife data coded inside that fall outside the domain. The C parameter is increased until a desired proportion of incorrectly classified data are correctly classified.

Distance Function Formalism

The first requirement in the calibration of distance function uncertainty is a dataset where all data locations have been coded as either inside or outside the domain:

$$i(\mathbf{u}_\alpha) = \begin{cases} 1, \text{ if domain present at } \mathbf{u}_\alpha \\ 0, \text{ otherwise} \end{cases}$$

For each sample located at $\mathbf{u}_\alpha$, the nearest sample location $\mathbf{u}_{\alpha'}$ is determined such that Min{ $\mathbf{u}_\alpha - \mathbf{u}_{\alpha'}$ }, $i(\mathbf{u}_\alpha) \neq i(\mathbf{u}_{\alpha'})$. The distance between the two locations is the distance function value at location $\mathbf{u}_\alpha$. If $\mathbf{u}_\alpha$ is within the domain, the distance is set to negative; otherwise the distance is positively signed:

$$df(\mathbf{u}_\alpha) = +(\mathbf{u}_\alpha - \mathbf{u}_{\alpha'}), \text{ if } i(\mathbf{u}_\alpha) = 0$$
$$df(\mathbf{u}_\alpha) = -(\mathbf{u}_\alpha - \mathbf{u}_{\alpha'}), \text{ if } i(\mathbf{u}_\alpha) = 1$$

The calculation of distance would account for anisotropy, if present. Notice that this approach is designed for binary systems where locations are in or out of a particular domain. Multiple domains could be modelled hierarchically. The presence of many intermingled domains would not be possible with this approach. Due to the nature of the distance function being negative inside the domain and positive outside the domain, a first guess for the bounding interface of interest would be the line or surface corresponding to a distance function value of zero. The distance function varies smoothly between increasingly positive values outside and further away from the boundary interface to increasingly negative values inside and further away from the boundary surface. Once distance function has been calculated for each sample, this distance function data can be interpolated on a regular grid using a smooth estimator such as kriging or inverse distance estimation. The boundary is considered to lay at the transition between positive and negative interpolated distance function values.

For example, consider the sample locations coded as inside and outside the domain in Figure 2a where black samples are inside. The distance to the nearest outside sample is calculated for each inside sample and vice versa. These distances are shown in Figure 2b. Samples outside the domain have positive distance function; samples inside the domain have negative distance function. These samples are interpolated yielding the map of values shown in Figure 2b. Negative estimates are considered inside the domain and positive estimates are considered outside the domain yielding the map shown in Figure 2c where black is inside the domain. An estimate of the boundary location falls at the transition between positive and negative distance function estimates.

C Parameter

This work proposes a new method for determining distance function uncertainty. This method uses the data to calibrate a single additive factor, *C*, which modifies the distance function values calculated at the sample locations. This method is similar to the jackknife where a subset of the data is held back when estimation is performed and estimated values are compared with the true values at sample locations.

The *C* parameter modifies the distance function value at each sample location. *C* is an additive parameter, being added to the distance function when outside the domain and subtracted from the distance function when inside the domain:

$$\hat{df}(\mathbf{u}_\alpha) = df(\mathbf{u}_\alpha) + C, \text{ if } i(\mathbf{u}_\alpha) = 0$$
$$\hat{df}(\mathbf{u}_\alpha) = df(\mathbf{u}_\alpha) - C, \text{ if } i(\mathbf{u}_\alpha) = 1$$

This modification is illustrated in Figure 3. The *C* parameter increases the difference between positive and negative distance function values. Once the *C* parameter has been applied to the data, the modified distance

function is interpolated. Modified distance function estimates greater than *C* are considered outside the domain. Modified distance function estimates less than *–C* are considered inside the domain. Any modified distance function estimates between *–C* and *C* are within the range of boundary uncertainty; the boundary is located between modified distance function estimates of *–C* and *C*.

The uncertainty band for different *C* values is illustrated in Figure 4 for different values of *C*. The same data shown in Figure 2 are used. A *C* value of 4 increases the positive and decreases the negative distance function values by 4. The modified distance function values are interpolated. Any modified distance function estimate greater than 4 is considered outside the domain (white) while any modified distance function estimate less than -4 is considered inside the domain (black). The grey areas have a modified distance function estimate between -4 and 4 and represent the region within which the boundary may lay. This is repeated for *C* values of 6 and 8. As *C* increases, the data values change and the size of the grey boundary uncertainty region increases. There is a need to infer a reasonable *C* value for each boundary.

C Calibration

The *C* parameter controlling the size of the boundary uncertainty is calibrated in a manner similar to the jackknife. A subset of the data is removed and the remaining data are used to estimate distance function at the jackknife locations. The number of jackknife data that fall on the wrong side of the boundary is reduced as *C* is increased. The first step in the calibration of *C* is to remove a subset of the data. This can be done by randomly choosing drillholes to exclude. Distance function values are then calculated for the remaining data with an initial value of *C*=0. The tool `DFCalc` performs these operations. The user specifies the proportion of drillholes they wish to be removed. They also specify the *C* parameter. This should initially be set to zero. Two files are written out. One contains the coordinates and distance function values for the data not excluded, the other contains the locations and domain flags for the excluded data.

Global kriging (Neufeld and Wilde, 2005) is then used to estimate distance function at each of the jackknife data locations. Global kriging is preferred as there are no artifacts due to the search for nearby data. All of the data are used to estimate at all locations. There are four possible outcomes resulting from this estimation as shown in Figure 5. The location could be: 1) correctly estimated to be outside the domain, 2) correctly estimated to be inside the domain, 3) incorrectly estimated to be outside the domain, and 4) incorrectly estimated to be inside the domain. We are interested in the number of data that fall on the wrong side of the boundary, that is, the number of times the estimate is positive but the data is coded as inside the domain and the number of times the estimate is negative but the data is coded as outside the domain. The number of times a data falls on the wrong side of the boundary for *C*=0 is the base case.

C is then increased and the modified distance function values at sample locations calculated. Modified distance function is estimated at each of the jackknife locations. The boundary is now considered to fall between -*C* and *C*. A data falls on the wrong side of the boundary when either a jackknife location is coded as inside but has a modified distance function estimate greater than *C* or a jackknife location is coded as outside but has a modified distance function estimate less than *–C*. The number of data falling on the wrong side of the boundary decreases as C increases similar to the manner shown in Figure 6. C is increased until the number of data falling on the wrong side of the boundary is acceptable. The *C* value where this occurs is the calibrated *C* value which describes the boundary uncertainty. The tool `GJCC` (stands for Global Jackknife *C* Calibration) checks the number of data on the wrong side of the boundary for a given *C*.

This is illustrated in Figure 7 using the same data as previously. The jackknife data not used in the initial estimation of distance function are considered. The white-filled circles represent sample locations outside the domain while the black-filled circles represent sample locations inside the domain. For *C*=0, there are two samples coded as outside the domain which fall inside (white circles in black region) and four samples coded as inside the domain which fall outside (black circles in white region) for a base case of six incorrectly classified data. Increasing

the *C* parameter to 4 decreases the number of samples coded as outside the domain which fall inside from two to one (one white bullet that was in the black region now falls in the grey region) and decreases the number of samples coded as inside the domain which fall outside from four to three (one black bullet that was in the white region now falls in the grey region). Increasing *C* to 6 further decreases the number of black circles falling in the white region to two for a total of 3 incorrectly classified data, one half of the base case at *C*=0.

Once *C* has been determined it is used to calculate the modified distance function at all sample locations. The *C* parameter is applied as shown previously; it is added to positive distance function values and subtracted from negative distance function values. The resulting modified distance function is then interpolated using global kriging. The boundary is considered to fall inside the distance function transition from -*C* to *C*. Different boundaries can be extracted by applying a threshold between -*C* and *C*. Figure 8 shows an example of three different thresholds applied to arrive at three boundaries: dilated, median and eroded. The dilated case is big everywhere and the eroded case is small everywhere.

Conclusions

Determining stationary domains is an important step in geostatistical modeling. The locations of the boundaries between stationary domains must be determined. There is uncertainty in the locations of the boundaries between stationary domains away from data. Uncertainty in boundary locations should be accounted for. The uncertainty should be fair. The boundary uncertainty can be calibrated using the available data in a manner similar to the jackknife. A subset of data is removed and the distance function is calculated at the remaining sample locations. Distance function is then estimated at the jackknife locations. A number of jackknife locations fall on the wrong side of the boundary. This quantity is reduced as increasing boundary uncertainty is considered by increasing the additive *C* parameter. The *C* value is calibrated by considering increasing values of *C*. The value of *C* which sufficiently reduces the number of jackknife data on the wrong side of the boundary summarizes boundary uncertainty.

References

Hosseini, A.H., 2009, Probabilistic Modeling of Natural Attenuation of Petroleum Hydrocarbons, *Ph.D. Thesis*, University of Alberta, 359p.

McLennan, J., 2008. The Decision of Stationarity, *Ph.D. Thesis*, University of Alberta, 191p.

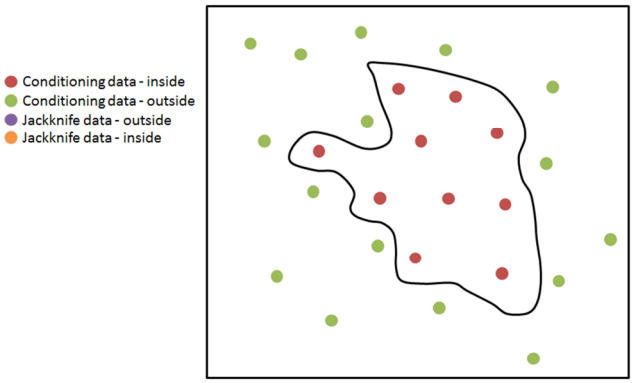
Munroe, M.J. and Deutsch, C.V., 2008a. A methodology for modeling vein-type deposit tonnage uncertainty. Center for Computational Geostatistics Annual Report 10. University of Alberta. 10p.

Munroe, M.J. and Deutsch, C.V., 2008b. Full calibration of *C* and β in the framework of vein-type deposit tonnage uncertainty. Center for Computational Geostatistics Annual Report 10. University of Alberta. 16p.

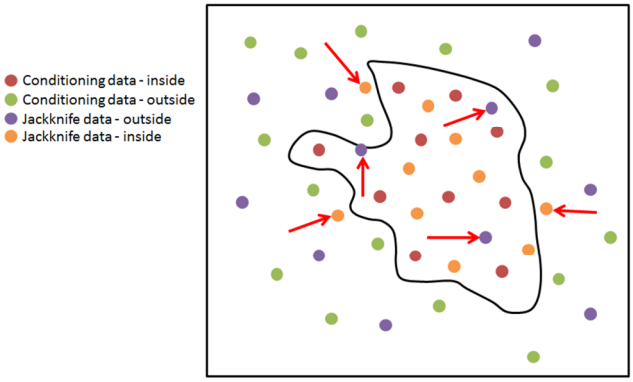
Neufeld, C.T. and Wilde, B.J., 2005. A global kriging program for artifact-free maps. Center for Computational Geostatistics Annual Report 7. University of Alberta. 8 p.

Figures

1. Interpolate distance function using a data subset:



2. Check the number of jackknife data that fall on the wrong side of the boundary:



3. Increase the *C* parameter until the number of data falling on the wrong side of the boundary is acceptable:

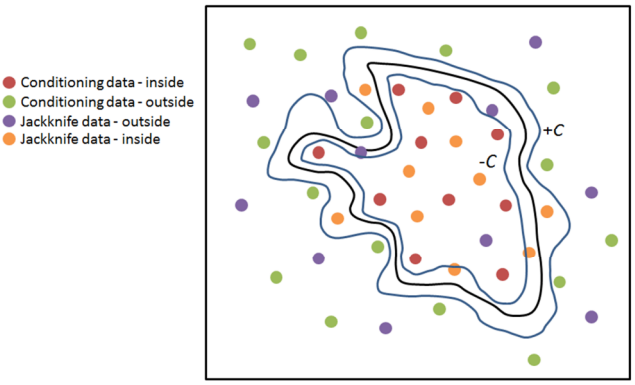


Figure 1: Illustration of data-based calibration of *C* parameter.

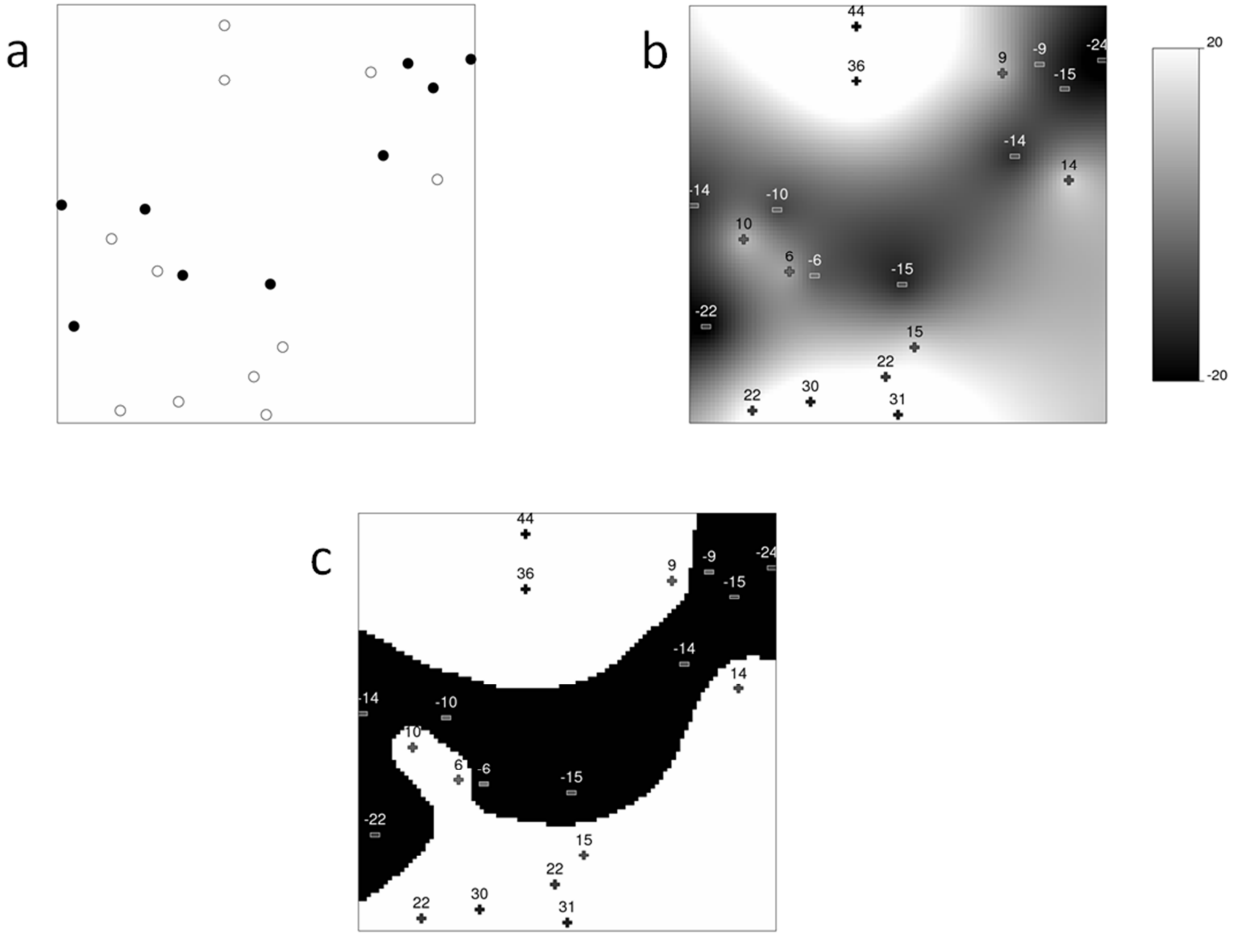


Figure 2: a) Locations coded as inside (black) and outside (white) the domain, b) distance function calculated at each sample location and interpolated, c) distance function less than zero considered inside the domain.

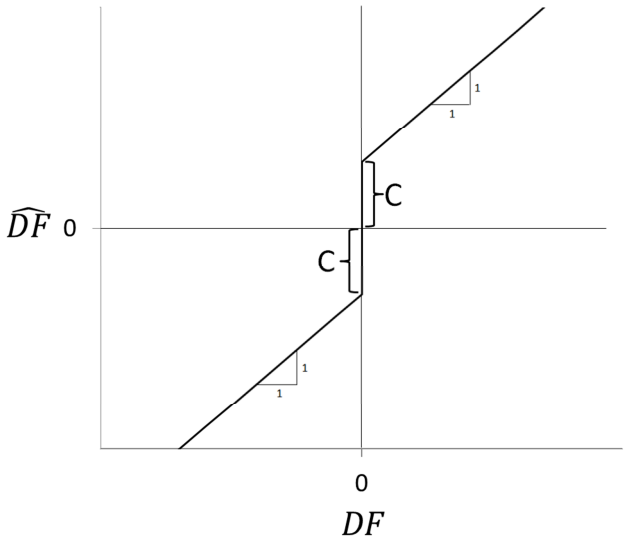


Figure 3: Conversion of distance function to modified distance function by the *C* parameter.

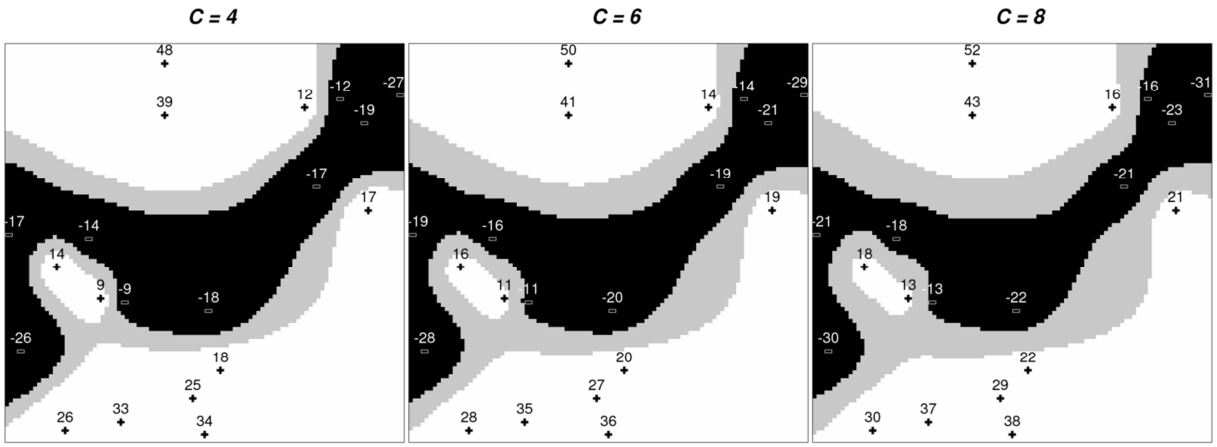


Figure 4: Boundary uncertainty for different values of *C*. Black=surely inside domain, white=surely outside domain, grey=region of boundary uncertainty.

		Global Kriging Estimate	
		Positive	**Negative**
Jackknife Code	**In**	**Estimated Out Truly In**	**Correctly In**
	Out	**Correctly Out**	**Esimated In Truly Out**

Figure 5: Possible outcomes for distance function estimation at jackknife data location.

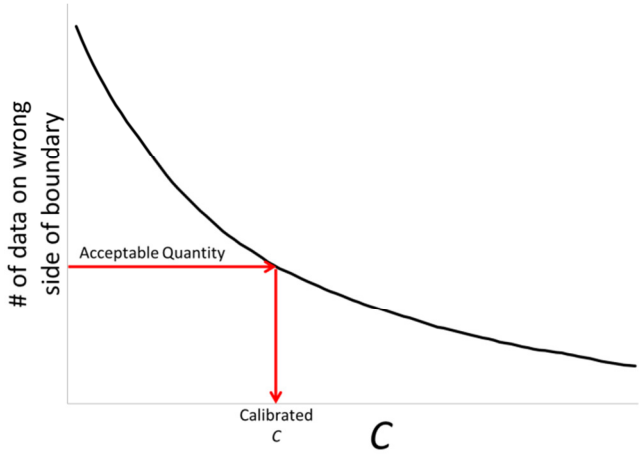


Figure 6: Idealized relationship between the number of data on the wrong side of the boundary and *C*.

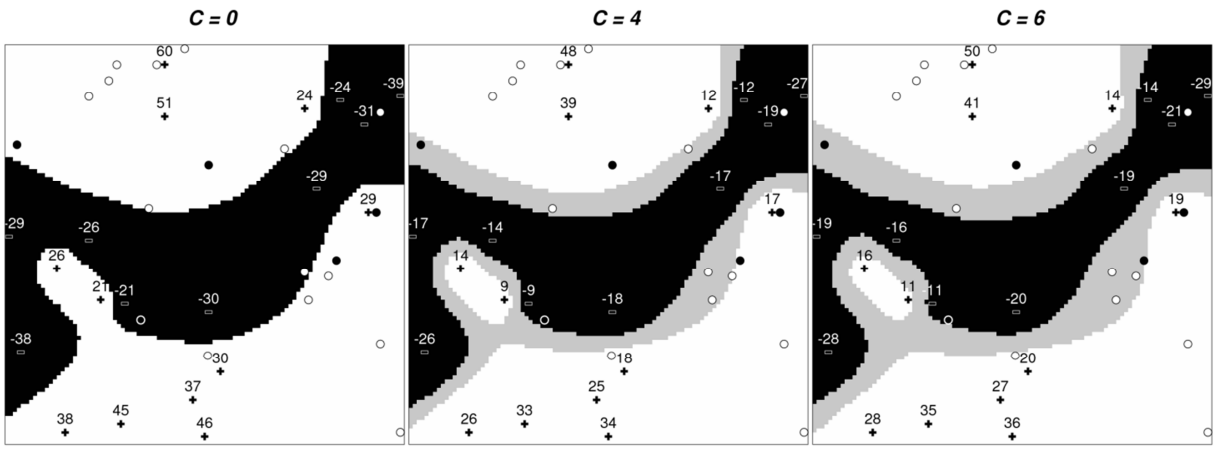


Figure 7: Jackknife data (black and white circles) are used to calibrate *C*.

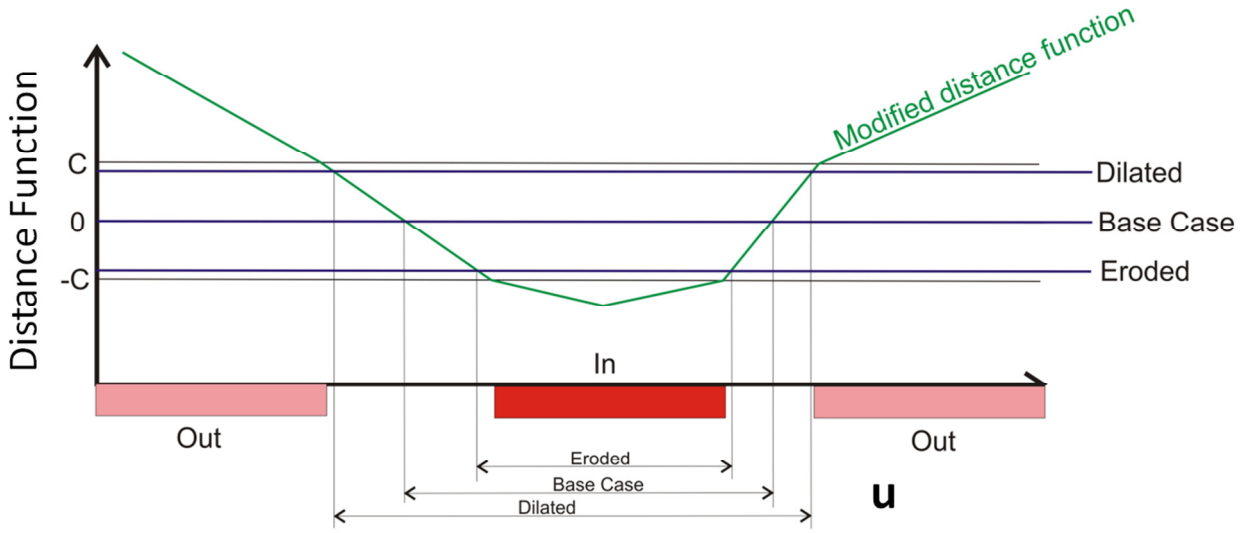


Figure 8: 1D schematic of distance function thresholds applied to arrive at different boundary locations.