# A Hybrid Approach to Model Selection for Support Vector Classification for Categorical Data

Enrique Gallardo and Oy Leuangthong

*The support vector machine – SVM algorithm is a Learning Machine that has been successfully applied to the problem of estimating the value of categorical data (e.g. facies or rock types) at unsampled locations. Roughly speaking, the algorithm looks for a linear boundary that balances between separating the observed facies with a maximum margin and to classify correctly all of them. The boundary calculated by this approach is expected to have a good generalization property and it is used to assign the facies at unsampled locations. The SVM algorithm implemented with a Gaussian radial basis function kernel requires determining an optimal combination of two parameters, the penalty C of the optimization problem and the γ of the kernel. Typically, these parameters are selected by cross-validation (e.g. LOO, K-fold) on the observed data. This paper proposes a hybrid technique in which the parameters are selected based on the average classification accuracy of the SVM algorithm over a set of geostatistical realizations for categorical data. Unlike conventional model selection approaches, where only the observed data is used through the cross-validation technique to obtain an estimated of the generalization accuracy, the proposed method interprets the set of geostatistical realizations as equally probable representations of the target true set to be classified. The realizations are used to obtain a new proxy of the generalization accuracy. Both conventional cross validation and this hybrid technique select the parameters of the SVM that leads to the maximum generalization accuracy estimate. A synthetic application shows that the hybrid approach to model selection produces comparable or better results than conventional cross validation in terms of generalization accuracy and robustness. These properties lead to more accurate and reliable prediction models.*

## Introduction

Support Vector Machines (SVM) has been successfully applied to subsurface geological characterization (e.g. Kanevski et al., 2001; Wohlberg, Tartakovky and Guadagnini, 2006). The characterization of geological sites is similar to solving the classification problem of assigning a single category (facies or rock types) to unsampled locations based on observed data. Implementing SVM requires the simultaneous selection of two parameters that define the SVM model and determine its performance: the penalty parameter C and the kernel parameter. The task of selecting these parameters is called model selection and it is the focus of this paper. A novel hybrid technique that makes use of geostatistical simulations to make model selection is proposed. The basic idea is interpreting a set of *L* geostatistical realizations as equally probable representations of the target true set to be classified. This interpretation allows using the realizations to calculate *L* estimates of the generalization accuracy for each SVM model on a predefined grid-search of parameter values. The SVM model with the highest mean generalization accuracy estimate is selected.

## Model selection and generalization accuracy estimate

Implementing SVM for classification problems (also known as Support Vector Classification – SVC) requires the simultaneous selection of the penalty (*C*) and the kernel parameters. The parameter *C* appears in the mathematical formulation of the SVM algorithm, while the kernel parameter depends on the specific kernel selected. In practice, the linear, polynomial and the Gaussian radial basis function (Grbf) kernels are extensively used. The latter, implemented in this research, has the form:

$$K(\mathbf{u},\mathbf{u}') = \exp\left(-\gamma\left\|\mathbf{u}-\mathbf{u}'\right\|^2\right);\ \gamma > 0$$

where **u** and **u'** represent any two different locations in the domain under study, and γ is the parameter that must be selected. A detailed description of SVM and/or kernels is not offered in this paper. The reader is referred to Vapnik (1998), Burges (1998) and Cristianini and Shawe-Taylor (2001).

The pair of parameters (C, γ) defines the SVM model along with its performance, which is measured through the generalization accuracy. Precisely, the purpose of model selection is to find the optimal SVM model (or the optimal pair (C, γ)) with the highest generalization accuracy. In practice, it is impossible to

know the "true" generalization accuracy; in consequence, an estimated value is used to select the optimal SVM model. Clearly, a technique that produces good estimates of the generalization accuracy is a technique that makes good selection of the parameters *C* and γ.

**Hybrid SVM model selection using Geostatistical simulation**

If the category (facies or rock types) at every location in the domain under study were known, the best SVM model would be found by calculating the true generalization accuracy; of course, the model that correctly classified the larger number of locations would be optimal.

This is not a real scenario; the true category at every unsampled location is unknown. However, geostatistical simulation algorithms can generate multiple realizations that populate with a single category the unsampled locations in the domain of interest. These realizations are equally probable and reproduce the sampled set and the spatial structure of the data. Since the realizations are reasonably valid descriptions of the unknown truth, they can be used to calculate multiple estimates of the generalization accuracy for each SVM model. The SVM model with the highest mean estimated generalization accuracy is selected. Describing the algorithms for generating geostatistical simulated realization is beyond the scope of this paper, the interested reader in geostatistics and its simulation algorithms is referred to the books of Goovaerts (1997), Chiles and Delfiner (1999) and Deutsch (2002).

The proposed technique is illustrated in figure 1. After selecting a searching space for the pair of parameters (C,γ), the procedure is:

a) Train and test the SVM at each node of the space of parameters *C* and γ.
b) Compare location-to-location the SVM response to each *L* simulated geostatistical realizations to calculate *L* estimates of the generalization accuracy. The simulated realizations must be previously generated using a technique suitable for the given set of data.
c) Calculate the mean of the *L* generalization accuracy estimates.
d) Generate a contour map or a surface of the mean generalization accuracy estimates. Select the SVM model (the pair (C,γ)) with the highest value.

**Leave-one-out (LOO)**

The classical cross-validation LOO method is applied in this paper for comparison purposes. The description of LOO along with a benchmark with others techniques for model selection can be found in Anguita, Boni, Ridella, Rivieccio and Sterpi (2005).

To illustrate the proposed methodology, a synthetic reference 2D data set of two facies was generated and 6 subsets of different sizes were sampled to be used as observed or training data. The proposed technique and LOO were analyzed.

The synthetic reference data set was generated by unconditional Gaussian simulation using an isotropic spherical semivariogram model without nugget effect and a range of 200 m. The Gaussian field was transformed to a categorical variable of two facies, white and black, using a threshold of 0 normal units. The small scale variability in the resulting map was removed by applying twice a moving window cleaning procedure. The reference data (Figure 2) has a resolution of 100 m x 100 m spacing that spans an area of 1km x 1km, so it contains 10000 nodes.

Six samples of different sizes (25, 50, 75,100,125 and 150 locations) were randomly drawn from the reference map to be used as training sets; in each case, the remaining unsampled locations were used as test set. The location maps of the sampled data are shown in appendix A.

To apply the proposed technique, for each set of sampled data 100 realizations were generated using the sequential indicator -SIS algorithm (Journel and Alabert, 1988; Alabert and Massonat, 1990; Deutsch and Journel, 1998). The noise in the geostatistical realizations was removed using the maximum a posteriori selection technique (MAPS) (Deutsch, 1998).

The searching space for the parameters C and *γ* is a 81x81 grid defined for the sequence of values C= $\{2^{-2}, 2^{-1.9},\ldots, 2^{5.9}, 2^{6}\}$ and γ= $\{2^{-2}, 2^{-1.9},\ldots,2^{5.9}, 2^{6}\}$. For each pair (C, γ) in the grid, an estimation of the generalization accuracy is calculated using the proposed technique and LOO. For each technique, the pair (C, γ) with the highest generalization accuracy estimate is used to classify all the unsampled locations in the test set. The classified locations allow calculating the generalization accuracy on the reference data (or "true" generalization accuracy). Additionally, for comparison purposes, the reference data was used to find the

SVC model with the best possible generalization accuracy estimate (which in this case is equal to the "true" generalization accuracy).

The GSLIB (Deutsch and Journel, 1998) and LIBSVM (Chang and Lin, 2001) software were used to make the geostatistical and SVC tasks presented in this paper, respectively.

**Analysis of results**

The proposed technique and LOO were analyzed and compared on their ability to: (1) find the optimal pair (C, γ) that leads to the SVC model with the highest "true" generalization accuracy; and (2) generate a good estimate of that value. The results are summarized in Table 1 and Figures 3, 4, 5 and 6.

The proposed hybrid technique out performed the LOO method in finding the optimal SVM model for almost all the sampled data sets (Figure 3). The new technique produces SVM models with better generalization accuracy than LOO. It is worth noting that the generalization accuracy obtained using the hybrid technique has similar values to those obtained using the reference data set. Figure 4 illustrates the results for the set of 50 samples. The reference data set and the proposed hybrid technique generate a smooth surface of the generalization accuracy estimates that allow selecting a unique optimum pair (C, γ). In contrast, the LOO technique generates an irregular surface with multiples steps that avoids selecting a unique optimum pair. These behaviors are consistent to all the training sets as it is evident in the contour maps plotted in appendix A.

**Table 1.** Model selection results and generalization accuracy estimates

| Number of samples | Reference (Test) | Estimated Hybrid (Mean) | Hybrid | Estimated LOO | Best LOO |
|---|---|---|---|---|---|
| 25 | 76,9 | 69,58 | 76,63 | 76 | 76,82 |
| 50 | 80,05 | 77,97 | 79,99 | 80 | 77,67 |
| 75 | 87,27 | 80,36 | 85,43 | 85,33 | 85,82 |
| 100 | 87,39 | 82,58 | 86,95 | 84 | 82,52 |
| 125 | 86,77 | 82,98 | 86,57 | 87,2 | 84,64 |
| 150 | 90,3 | 85,42 | 89,32 | 90 | 88,56 |

In terms of the ability to produce good estimates of the generalization accuracy, the hybrid technique provides values that are reasonably closer to the generalization accuracy calculated on the reference data. It is interesting to note that the hybrid technique never overestimated the "true" generalization accuracy value. Additionally, it allows calculating a range of values for the estimated generalization accuracy. The generalization accuracy estimates obtained by LOO tends to be too optimistic (Figure 6). Moreover, for small data sets ( < 100 data ) and the selected grid-search, LOO was unable to provide a unique optimal SVM model. Figure 4 shows that the surface map of the LOO generalization accuracy estimates is irregular and exhibits multiple areas with the same maximum value. Due to the difficulty in selecting a unique SVM model, Figure 6 shows the best and worst LOO generalization accuracy values calculated on the reference data.

**Conclusions**

A novel technique to make SVC model selection was introduced. The proposed hybrid technique interprets *L* geostatistical simulated realizations as equally probable representations of the unknown reality to calculate *L* estimates of the generalization accuracy for each pair of SVC parameters on a preselected grid-search. The pair that generates the SVC model with the highest mean generalization accuracy estimate is selected. A range of minimum and maximum values for this estimate can also be calculated. The study case shows that the hybrid technique produces good generalization accuracy estimates; in fact, the results were very close to the values obtained using the reference data. Further, the hybrid technique out

performed the classical LOO method in model selection and estimating the generalization accuracy. These results suggest that the hybrid technique is a good method for model selection.

**References**

1. Alabert, F.G., and Massonnat G.J. (1990) Heterogeneity in a complex turbiditic reservoir: Stochastic modelling of facies and petrophysical variability. In 65th Annual Technical Conference and Exhibition, pp. 775–790. SPE paper # 20604.
2. Anguita, D., Boni, A., Ridella, S., Rivieccio, F., and Sterpi, D. (2005) Theoretical and practical model selection methods for support vector classifiers, StudFuzz 177, pp. 159–179. Springer-Verlag.
3. Burges, C.J. (1998) A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2, pp. 121-167. Kluwer Academic Publishers, Boston.
4. Chiles, J.P., and Delfiner, P. (1999) Geostatistics: modeling spatial uncertainty. Wiley- Interscience, New York.
5. Chang, C.-C. and C.-J.Lin. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Retrieved 22 December 2008.
6. Cristianini, N., and Shawe–Taylor, J. (2001) An introduction to Support Vector Machines. Cambridge University Press.
7. Deutsch, C.V., and Journel, A.G. (1998) GSLIB: geostatistical software library and user's guide. Oxford University Press, New York.
8. Deutsch. C.V. (1998) Cleaning categorical variable (lithofacies) realizations with maximum a-posteriori selection. Computers & Geosciences Vol.24, No. 6, pp. 551-562.
9. Goovaerts, P. (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York.
10. Journel, A., and Alabert, F. (1988) Focusing on spatial connectivity of extreme valued attributes: stochastic indicator models of reservoir heterogeneities. SPE paper # 18324.
11. Kanevski, M., Canu, S., Maignan, M., Wong, P., Pozdnukhov, A., Shibli, S. (2001) Support vector machines for classification and mapping of reservoir data. IDIAP Research Report. IDIAP-RR-01-04.
12. Vapnik, V. (1998) Statistical learning theory. Wiley, New York
13. Wohlberg, B., Tartakovsky, D.M., and Guadagnini, A. (2006) Subsurface characterization with support vector machines. IEEE transactions on geoscience and remote sensing, Vol. 44, No. 1, Jan. 2006.

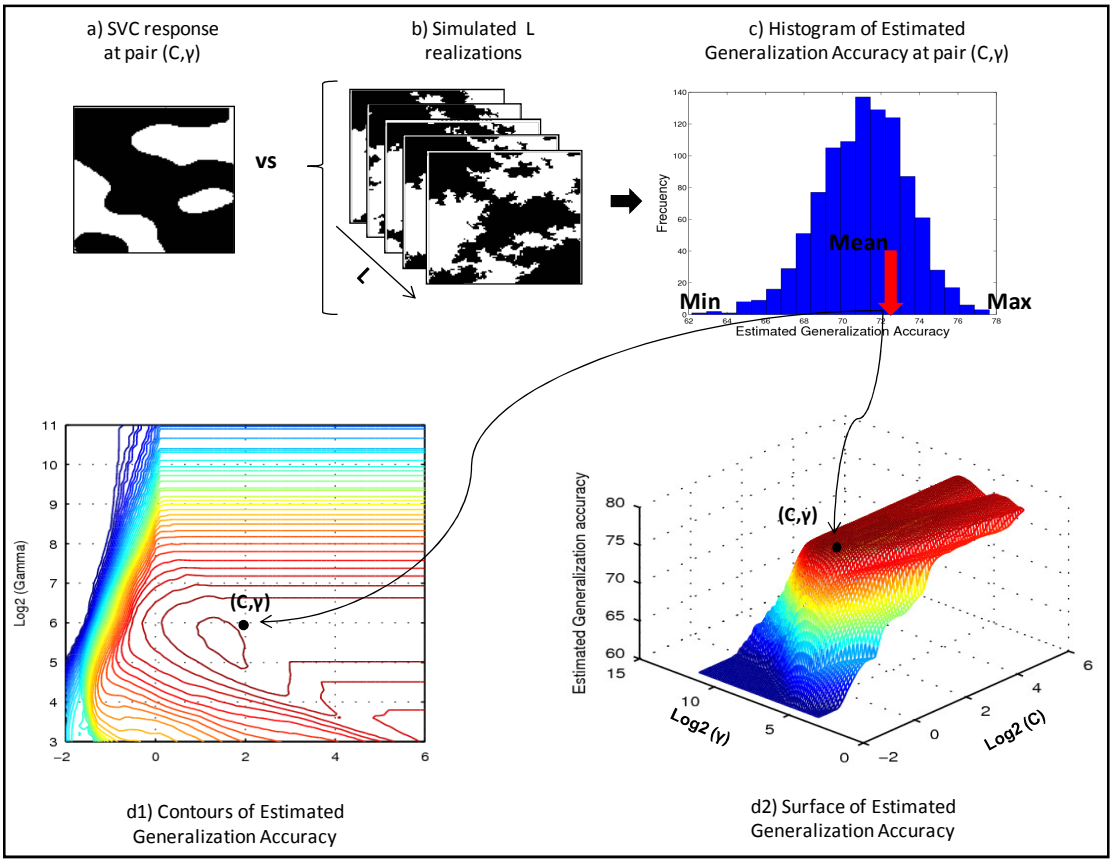



**Figure 1:** Illustration of the proposed hybrid technique

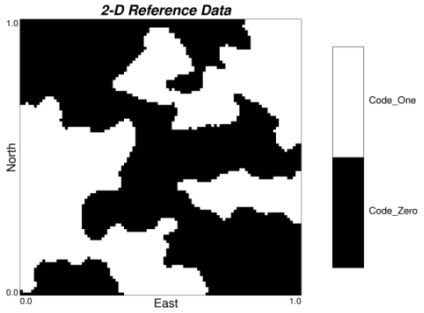



**Figure 2:** Synthetic reference data

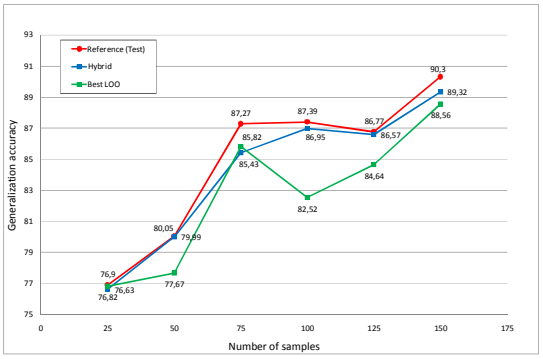



**Figure 3:** Model selection results. Lines illustrate the generalization accuracy calculated on the reference data set using the parameters determined by the different methods.

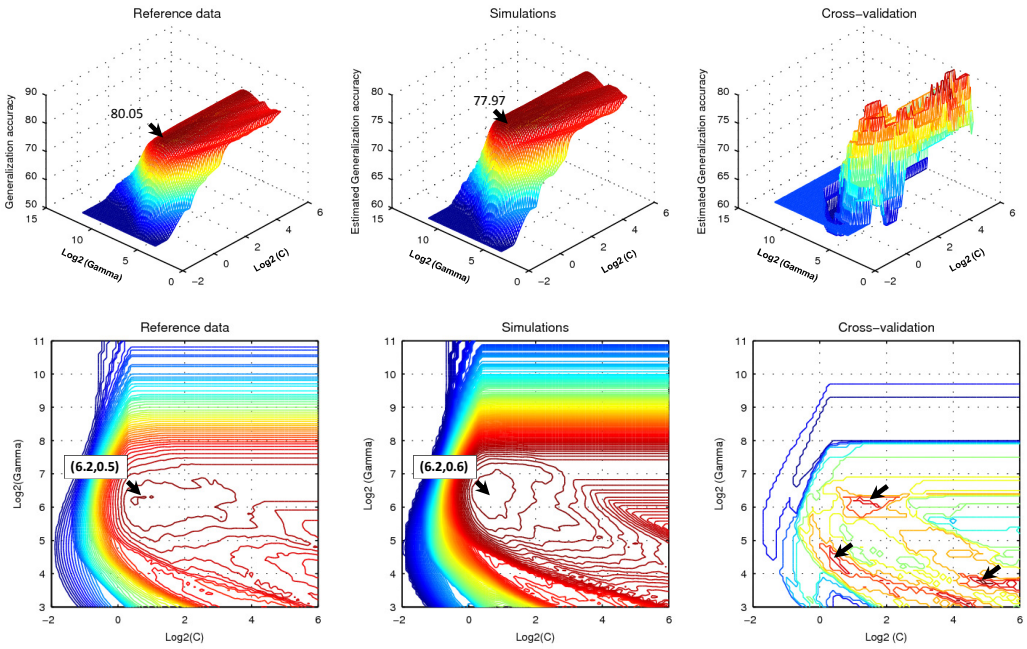



**Figure 4**: Detailed results for the set of 50 samples. Surfaces of the mean generalization accuracy estimates (top line) and contour lines of the mean generalization accuracy estimates (bottom line).

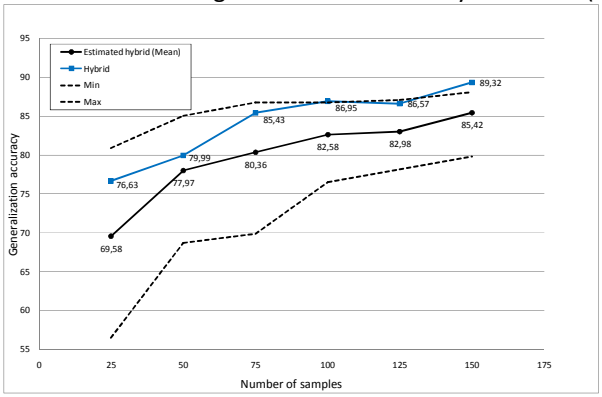



**Figure 5:** Results of proposed methodology. Black line shows the highest mean generalization accuracy estimates. Dashed lines show the maximun and minimun generalization accuracyestimates. Blue line shows the generalization accuracy calculated on the reference data.

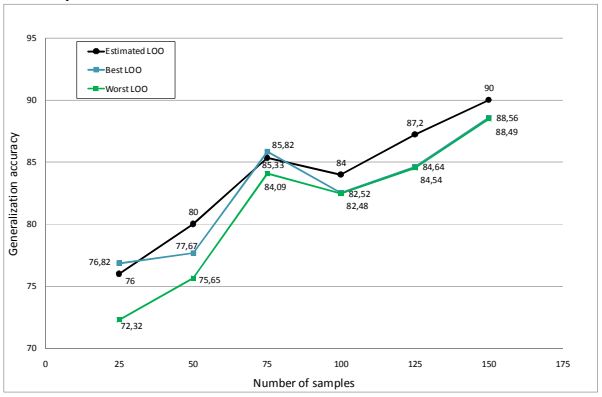



**Figure 6:** Results of LOO method. Black line shows the highest generalization accuracy estimates. Blue and green lines shows the best and worst generalization accuracy calculated on the reference data.

# APPENDIX A

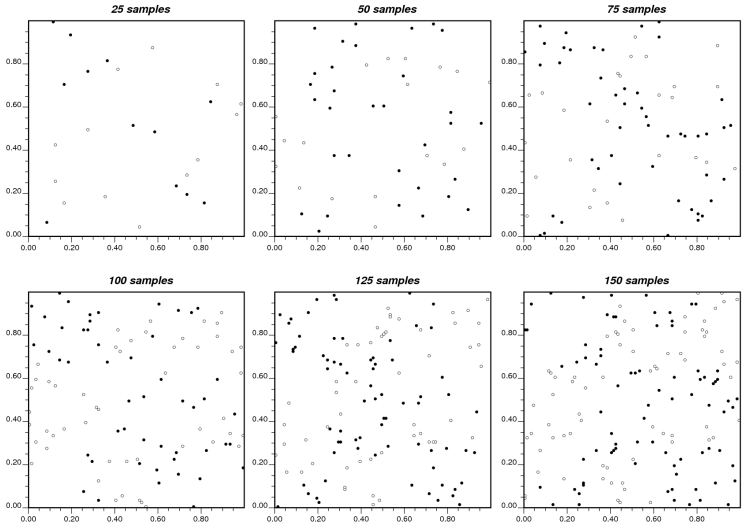
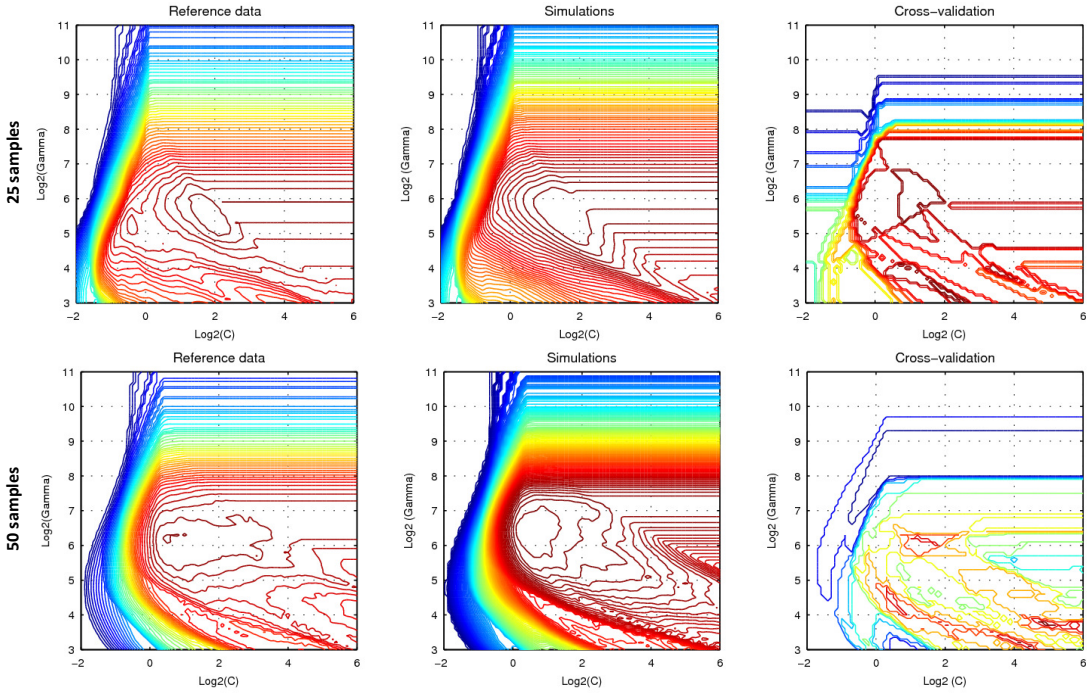



**Figure A1.** Location maps of the sampled locations (training sets)



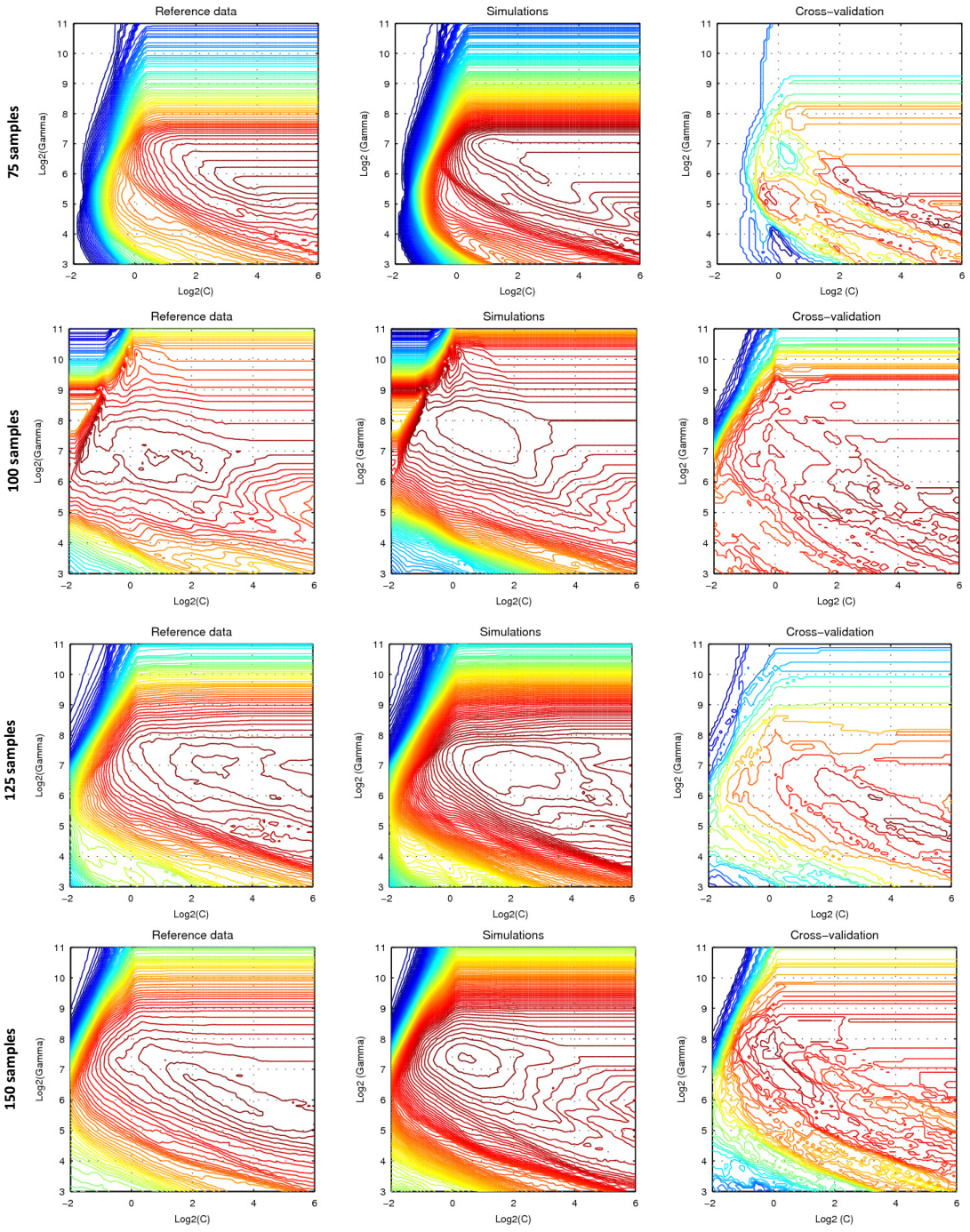



**Figure A2**. Contour plots of the generalization accuracy for each sampled size and each model selection method. The first column shows the results obtained using the reference data, the second column the results obtained using the proposed technique and the third column the results using LOO cross-validation.