Extraction of Representative Learning Set from Measured Geospatial Data

Béla Paláncz

Department of Photogrammetry and Geoinformatics, Faculty of Civil Engineering, Budapest University of Technology and Economics
Műegyetem rkp. 3, H-1111 Budapest, Hungary, palancz@epito.bme.hu

Lajos Völgyesi, Piroska Zaletnyik

Department of Geodesy and Surveying, Faculty of Civil Engineering, Budapest University of Technology and Economics
lvolgyesi@epito.bme.hu

Levente Kovács

Department of Control Engineering and Information Technology, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics
lkovacs@iit.bme.hu

Abstract: The efficiency of the application of soft computing methods like Artificial Neural Networks (ANN) or Support Vector Machines (SVM) depends considerably on the representativeness of the learning sample set employed for training the model. In this study a simple method based on the Coefficient of Representativity (CR) is proposed for extracting representative learning set from measured geospatial data. The method eliminating successively the sample points having low CR value from the dataset is implemented in Mathematica and its application is illlustrated by the data preparation for the correction model of the Hungarian gravimetrical geoid based on current GPS measurements.

Keywords: machine learing, representativness of data, geospatial data

1 Introduction

During the last decade, machine learning algorithms, such as artificial neural networks (ANN) and support vectors machines (SVM) have extensively used for wide range of applications. They have been applied for classification, regression, feature extraction, data prediction and spatial data analysis.

To ensure generalization properties of machine learning methods like artificial neural networks and support vector machines, the set of measured data should be split into learning and testing sets, [1]. The question is how to divide the measured sample set into these three sets in order to extract the most information as it is possible. This is especially important when the number of samples is relatively small. There are different methods suggested how to carry out the learning and testing process taking into account this requirement, [2]. Optimal sampling scheme would be regular triangular or square grids, which keep the maximum standard error to a minimum, [3]. However, geospatial data samples are irregularly spaced and do not form rectangular grid. Qualitatively these irregularities are indicated by local clustering and dispersion, but for numerical computations one needs quantitative characterization of the deviation from the optimal, uniform spatial sample distribution. There are different indices introduced to indicate the representativeness of a real sample distribution, [4]. In this study we employed the Coefficient of Representativity (CR) proposed by [4].

2 Measures of Representativity

Let us suppose, that we have $\{x_i, y_i, z_i\}$ measured sample points and their $\{x_i, y_i\}$ coordinates are on a convex region, see Figure 1.

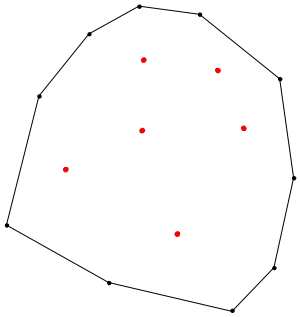

Figure 1
Measured data sample points and the border of the convex region

2.1 Nearest Neighbours Index

One of the possible characterizations of the representativity of this sample set was suggested by [5] via *Nearest Neighbours Index* (*NNI*). The *NNI* is defined as the ratio of the mean of the *Nearest Neigbours distances* (NNI_{dist}):

$$MeanNN_{dist} = \sum_{i=1}^{N} \frac{NN_{dist}}{N} \tag{1}$$

where N is the number of sampling points and to the mean of the *Nearest Neigbours distances* for uniform distribution of the points. This *Mean Random Distance* (*MRD*) is defined as:

$$MeanRD = \sqrt{\frac{S_{Total}}{N}} \tag{2}$$

where S_{Toral} is the total surface of the investigated region. Thus the *NNI* is equal to:

$$NNI = \frac{MeanNN_{dist}}{MeanRD} \tag{3}$$

The *NNI* is close to 1 for the sampling points having a uniform spatial distribution. When $NNI < 1$, the samples are more clustered than expected compared to a uniform random distribution. In the contrary, an $NNI > 1$ indicates a dispersion of the samples.

The main limitation of this index is that this is a global measure, and gives no information about local clusters or dispersions.

2.2 Voronoi Polygons

Voronoi polygons have the property to contain only one measurement and to have a geometry that will include all the datapoints that are closer to the measurement than those associated to clustered data, [6]. The area of the Voronoi polygon belonging to a sample point may be considered as the region of attraction of this point, because the points of this region are closer to this sample points than to other sample points, see Figure 2.

In case of uniform distribution of the sample points, the size of the region of attraction of every sample point – the ares of the corresponding Voronoi polygons – is the same.

Therefore the histogram of the areas of these polygons might help describe quantitatively the homogenity of the sample set.

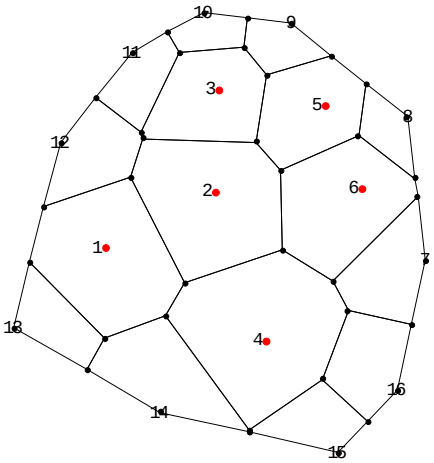

Figure 2
Voronoi polygons of the data samples and the border points

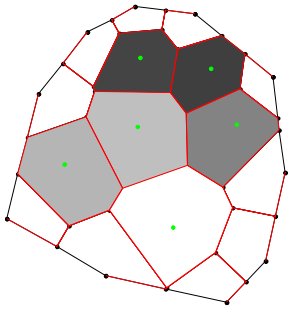

Figure 3
Intensity plot of the Voronoi polygons corresponding to their size

Figure 3 shows the Voronoi polygons, where a polygon gray level intensity is proportional with its size. Larger polygons are brigthter.

The main handicape of this measure is that points can be clustered and still have relatively large Voronoi polygons. In an other words, large Voronoi polygons do not guarantee that the points are isolated.

For example, the Voronoi polygon belonging to point 6 is larger than those belonging to point 3 or point 5. However, the distance between points 3 - 5 is greater than the distance between points 5 - 6 (Figure 2).

2.3 Coefficient of Representativity

Dubois, [4], suggested a new measure that combines both the distance of each point to its nearest neigbour and the surface of the Voronois polygons. This measure, called *Coeffient of Representativity* (*CR*) is a product of two terms:

$$A = \frac{S_V}{S_m} \tag{4}$$

which will take into account the surface of the Voronoi polygon. It is equal to the ratio of the surface of the Voronoi polygon (S_V) to the ideal surface it should have to obtain in case of a homogeneous sample set. This surface is simply defined as the mean surface (S_m) that is the total area of the investigated region S_{Total}, divided by the number of sampling points N:

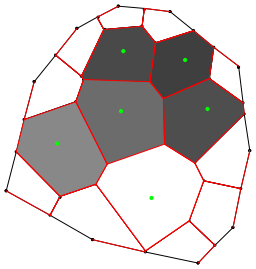

Figure 4
Intensity plot of the CR values. A polygon gray level intensity is proportional with its CR

$$S_m = \frac{S_{Total}}{N} \tag{5}$$

The second term B, is equal to the ratio of the squared distance between a point to its nearest neighbour (NN_{dist}) to the mean surface of the Voronoi polygons:

$$B = \frac{NN_{dist}^{\ 2}}{S_m} \tag{6}$$

For regular grid where points are distributed in the middle of each cell of grid $NN_{dist}^{\ 2} = S_V$ and $B = 1$. Then the CR for any point can be defined as:

$$CR = AB = \frac{S_V}{S_m}\frac{NN_{dist}^{\ 2}}{S_m} \tag{7}$$

Figure 4 shows the CR values of the Voronoi cells represented by gray level intensities. The measure based on the area of the Voronoi polygons are different from the measure based of CR, compare Figure 3 and Figure 4.

3 Constructing Optimal Learning Set

Once we have a measure of the representativity of a dataset, an algorithm can be developed to extract samples from the irregular dataset to form the best learning set as possible. This optimal extraction process can be considered as a combinatoric *max-min* problem. Namely, from the measured n patterns, one should select $m < n$ samples in a way, that in the constructed learning set the minimum of CR will be the greatest considering every possible $\binom{n}{m}$ combinations. Strictly saying, it is a $max(min(CR))$ combinatoric problem, and one may solve it by genetic algorithm.

However, such an algorithm is very time consuming, therefore a suboptimal algorithm may be employed as an alternative solution. In this case, we construct the learning set by eliminating sucessively samples from the original set of the n samples. Namely, we simply drop out the sample, which has actually the minimal CR and repeat this action $m - n$ times.

The implementation of this algorithm under *Mathematica* 5.2 is available in [8].

Let us eliminate two samples of the dataset, see Figure 1.

It can be clearly seen on Figure 5 comparing it with Figure 4, that the homogenity of sample set has been considerably impoved by elimination of the sample points having low CR values.

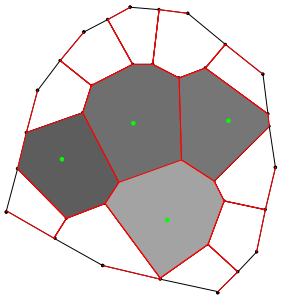

Figure 5
Intensity plot of the CR values after eliminating two samples

As illustration of the application of the method for real world problem, a learning set will be constructed for a neural network to be trained to model the Hungarian gravimetrical/GPS geoid.

4 Learning Set for the Hungarian Geoid

4.1 Data Preprocessing

Recently GPS measurements provide more precise data than gravimetrical measurements did before. However, their numbers are considerably less than those of the gravimetrical ones. Therefore it is reasonable to use them for correction. The values of the correction of the gravimetrical geoid –- the so-called corrector surface – are based on the differences between the GPS and the gravimetrical measurements, [7]. In case of Hungary we have the following dataset for the corrector surface, see Figure 6.

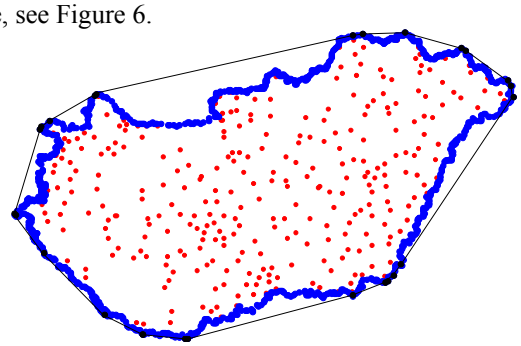

Figure 6
Locations of the sample values of the corrections and the convex border of the Hungarian region

Clustering and dispersion of the datapoints can be clearly seen on Figure 6.

4.2 Computing Voronoi Tesselations

First, we compute the Voronoi polygons, see Figure 7.

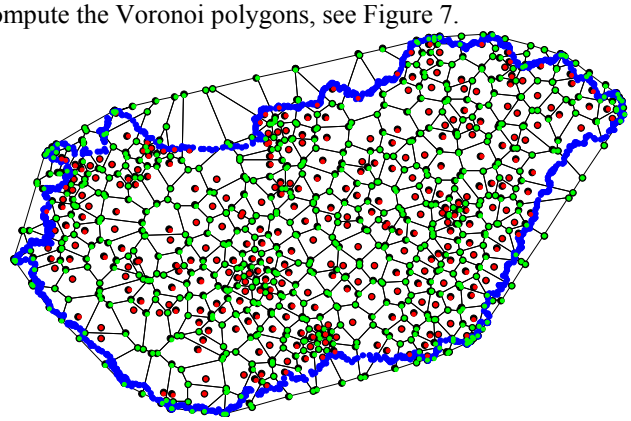

Figure 7
Voronoi tesselations

4.3 Computing Coefficient of Representativity

The CR values for the sample points can be computed, see Figure 8.

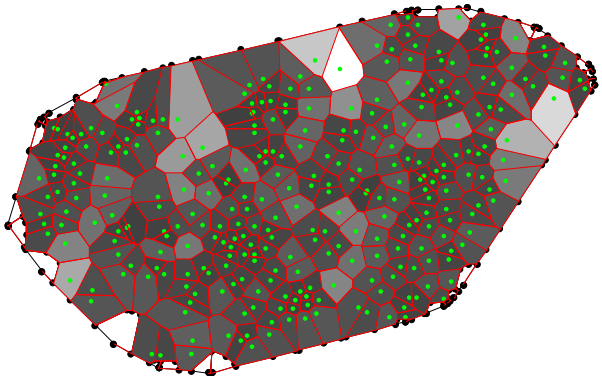

Figure 8
The distribution of CR in the Voronoi cells

Smaller the value of CR darker the corresponding cell region.

Figure 9 demonstrates the distribution of the CR, indicating the majority of the small values.

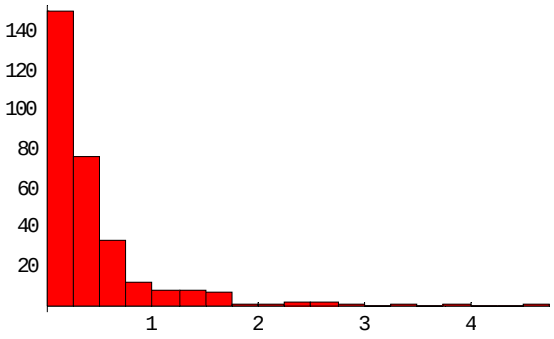

Figure 9
The histogram of the CR distribution of the original data set

The statistics of the CR distribution of the original sample set is showed in Tab. 1.

Table 1
Statistics of CR distribution of the original data set (304 points)

Min	Max	Mean	Standard deviation
0.00235	4.712	0.449	0.593

4.4 Sucessive Elimination of Sample Points Having Low CR

In order to create the learning set, we eliminate m = 110 sample points from the original n = 304 datapoints.

Figures 10-12 show the remained points after elimination as learning set, the Voronoi tessalation and the distribution of the CR values respectively.

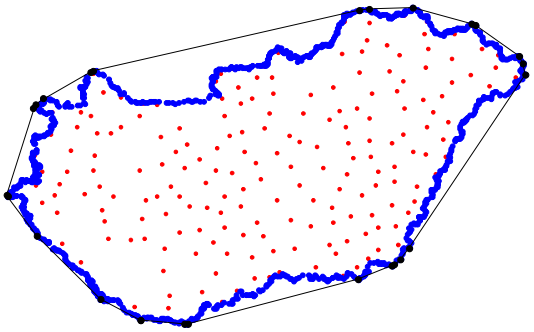

Figure 10
Locations of the sample values of the corrections after elimination of 110 points

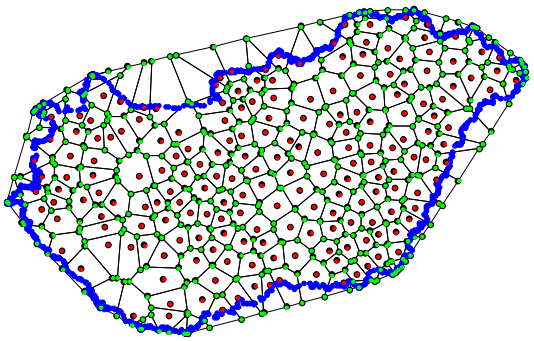

Figure 11
Voronoi tesselations of the learning set

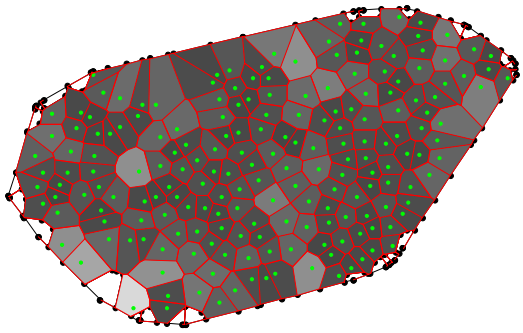

Figure 12
The distribution of CR in the Voronoi cells in the learning set

On Figure 13 can be seen how considerably changed the CR distribution.

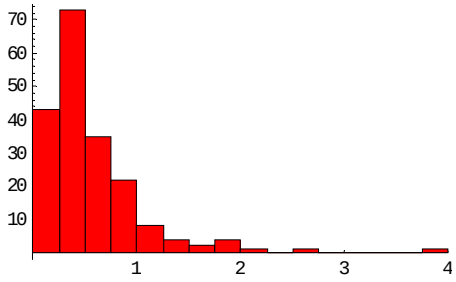

Figure 13
The histogram of the CR distribution in the learning set

The statistics of the CR distribution of the original sample set are in Table 2.

Table 2
Statistics of CR distribution of the learning set (194 points)

Min	Max	Mean	Standard deviation
0.1606	4.767	0.563	0.469

Conclusions

The suggested method is proved to be successful to decrease considerably the inhomogenity of the learning dataset and the differences in the CR indices of the data points. An improvement of this method would be the application of Voronoi tessalation on non-convex region. In this way the effect of non-convex country border can be taken into account and more realistic CR values could be computed.

Acknowledgement

The authors would like to thank A. Kenyeres providing the GPS/levelling data of Hungary.

References

[1] Berthold M., D. J. Hand (Eds.): Intelligent Data Analysis, An Introduction, Springer, 2003

[2] Gilardi N., S. Bengio: Local Machine Learning Models for Spatial Data Analysis, Journalof Geographic Information and Decision Analysis, 2000, Vol. 4/1, pp. 11-28

[3] McBratney A. B., R. Webster, T. M. Burgess: The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables. I. Theory and Method, Computer & Geosciences, 1981, Vol. 7/4, pp. 331-334

[4] Dubois G.: How Representative are Samples in Sampling Network?, Journal of Geographic Information and Decision Analysis, 2000, Vol. 4/1, pp. 1-10

[5] Clark P. J., F. C. Evans: Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations, Ecology, 1954, Vol. 35, pp. 445-453

[6] Okabe A., B. Boots, K. Sugihara: Spatial Tessellations. Concept and Applications of Voronoi Diagrams, Wiley and Sons, 1992

[7] Featherstone W. E.: Refinement of a Gravimetric Geoid Using GPS and Levelling Data, Journal of Surveying Engineering, 2000, Vol. 126/2, pp. 27-56

[8] Paláncz B., L. Völgyesi, P. Zaletnyik, L. Kovács: Computing Representative Learning Set via Mathematica, 2006, http://library.wolfram.com/infocenter/Mathsource