Correcting Order Relation Deviations for Categorical Variable

Jared L. Deutsch and Clayton V. Deutsch

Many geostatistical methods, such as indicator kriging inside sequential indicator simulation, introduce order relation deviations that must be corrected. The most common method for correcting these deviations unfairly treats categories with a small marginal probability. This leads to a bias in the results. The method implemented in this short note attempts to correct this by using the probability of each category and the probability of not *each category. Thus, categories with high probability and low probability are considered equally fairly. An initial implementation inside the BlockSIS program gives some promising results that warrant further investigation.*

Background

Consider *k*=1,…,*K* mutually exclusive and exhaustive categories. The probability of each category depends on location and relevant data (local hard data and soft secondary data). Many approaches to estimate the probability of each category, conditioned to all of the data, lead to results that do not satisfy the axioms of probability, that is, probabilities may not be non-negative and sum to one over the closed set of *K*. Consider a set of estimated probabilities or proportions p_k for *k*=1,…,*K*. The most common method for correcting order relation deviations is to set negative probabilities equal to 0 and rescale them to sum to one by dividing by the sum of all probabilities:

$$p_k = \frac{\left(p_k\right)_{\min=0}}{\sum_j \left(p_j\right)_{\min=0}} \tag{1}$$

This is problematic because high probabilities are treated differently from low probabilities introducing artifacts into the resulting distribution. As a simple example, consider probabilities of p_1=-0.1 and p_2=0.8. These would be corrected to 0.0 and 1.0, for *k*=1 and *k*=2 respectively; however, since the original p_2 is not *that* close to one, perhaps the corrected p_1 should be slightly above 0 and the corrected p_2 slightly below 1. The value of 0.8 is increased by 0.2 and the value of -0.1 is increased by only 0.1. Additive schemes to correct the probabilities are also problematic, because multiplicative rescaling is inevitably required. This short note presents another idea.

Rescaling Method

The key idea is the notion that 0=1, that is, the certain probability of something being true (1) is equivalent to the certain probability of something else being not true (0). To treat both low and high probabilities the same, we consider the complimentary probabilities for each category (Equations 2 and 3). These estimates are equally valid, so an equal weighted average is used to reconcile these disparate estimates (Equation 4). The resulting p_k^c satisfy order relations considering the probability of *k* and not *k* for the *K* binary evaluations; however they do not necessarily sum to unity. A final restandardization is applied to enforce this condition (Equation 5).

$$p_k^a = \frac{\left(p_k\right)_{\min=0}}{\left(p_k\right)_{\min=0} + \left(\sum_{j,j\neq k} p_j\right)_{\min=0}} \tag{2}$$

$$p_k^b = 1 - \frac{\left(1-p_k\right)_{\min=0}}{\left(1-p_k\right)_{\min=0} + \left(\sum_{j,j\neq k} \left(1-p_j\right)\right)_{\min=0}} \tag{3}$$

$$p_k^c = \frac{p_k^a + p_k^b}{2} \tag{4}$$

$$p_k^c = \frac{p_k^c}{\sum_j p_j^c} \tag{5}$$

For a more complete development of this rescaling procedure, the reader is encouraged to review Deutsch (2009) in this report.

Implementation and Testing with BlockSIS

The latest version of *BlockSIS* uses the conventional correction described in the Background. The new rescaling method reviewed here was implemented in *BlockSIS* as an alternative correction for order relation deviations. It has been suggested (Deutsch, 2005 and Ortiz, 2003) that order relation deviations are responsible for introducing a bias responsible for significant deviations in output proportions from input proportions. This effect can be especially significant for categories with low global proportions (5% or less).

To test the effect, an area measuring 100m x 100m x 5m with a grid cell size of 1m x 1m x 0.1m was unconditionally simulated in *BlockSIS* using simple kriging. Three categories: 0, 1 and 2 were given input global proportions of 0.05, 0.70 and 0.25 respectively. A number of variograms were tested using the conventional order relations correction until a variogram that produced a significant bias for Category 0 was created. The parameters and variogram specification for *BlockSIS* used are given below (Figure 1).

```
                  Parameters for BLOCKSIS
                  ***********************

START OF PARAMETERS:
0                              -0=SK,1=OK,2=L1,3=L2,4=CC,5=BU,6=PR,7=BK,8=BC
0                              -Clean: 0=none, 1=light, 2=heavy, 3=super
3                              -number of categories
0     1       2                -   categories
0.05  0.70    0.25             -   global proportions
0.50  0.50    0.50             -   correlation coefficients for soft data
nodata1                        -file with local data
1   2   3   4                  -   columns for X,Y,Z, and category
nodata2                        -file with gridded prior mean values
1   2   3                      -   columns for each category
3                              -   2-D areal map (2) or 3-D cube (3)
nodata3                        -file with keyout array
1                              -   column for keyout indicator
0                              -debugging level: 0,1,2,3,4
temp9                          -file for debugging output
temp8                          -file for simulation output
150                            -number of realizations
100 0.5 1.0                    -nx,xmn,xsiz
100 0.5 1.0                    -ny,ymn,ysiz
50  0.5 0.1                    -nz,zmn,zsiz
582073                         -random number seed
12                             -maximum original data  for each kriging
24                             -maximum previous nodes for each kriging
1                              -assign data to nodes? (0=no,1=yes)
0                              -maximum per octant    (0=not used)
25.0 25.0 25.0                 -maximum search radii
 0.0  0.0  0.0                 -angles for search ellipsoid
51   51   51                   -size of covariance lookup table
1 0.00                         -Cat 1: nst, nugget effect
1 1.00  0.0 0.0 0.0            -       it,cc,ang1,ang2,ang3
  20.0  20.0   2.0             -       a_hmax, a_hmin, a_vert
1 0.00                         -Cat 1: nst, nugget effect
1 1.00  0.0 0.0 0.0            -       it,cc,ang1,ang2,ang3
  60.0  60.0  20.0             -       a_hmax, a_hmin, a_vert
1 0.00                         -Cat 1: nst, nugget effect
1 1.00  0.0 0.0 0.0            -       it,cc,ang1,ang2,ang3
  60.0  60.0  20.0             -       a_hmax, a_hmin, a_vert
```

Figure 1: *BlockSIS* parameters with the conventional and test correction for order relation deviations

Using the conventional and proposed corrections, 150 realizations were generated using *BlockSIS*. Summary statistics of the simulation results are provided in Table 1. Histograms of the resulting output proportions are shown in Figure 2.

Table 1: Summary statistics for output from *BlockSIS* using the conventional and test correction

	Target Proportion	Conventional Mean Prop.	Test Mean Prop.	Conventional Std. Dev.	Test Std. Dev
Category 0	0.0500	0.0918	0.0554	0.0352	0.0282
Category 1	0.7000	0.6508	0.6619	0.1459	0.1438
Category 2	0.2500	0.2574	0.2827	0.1479	0.1398

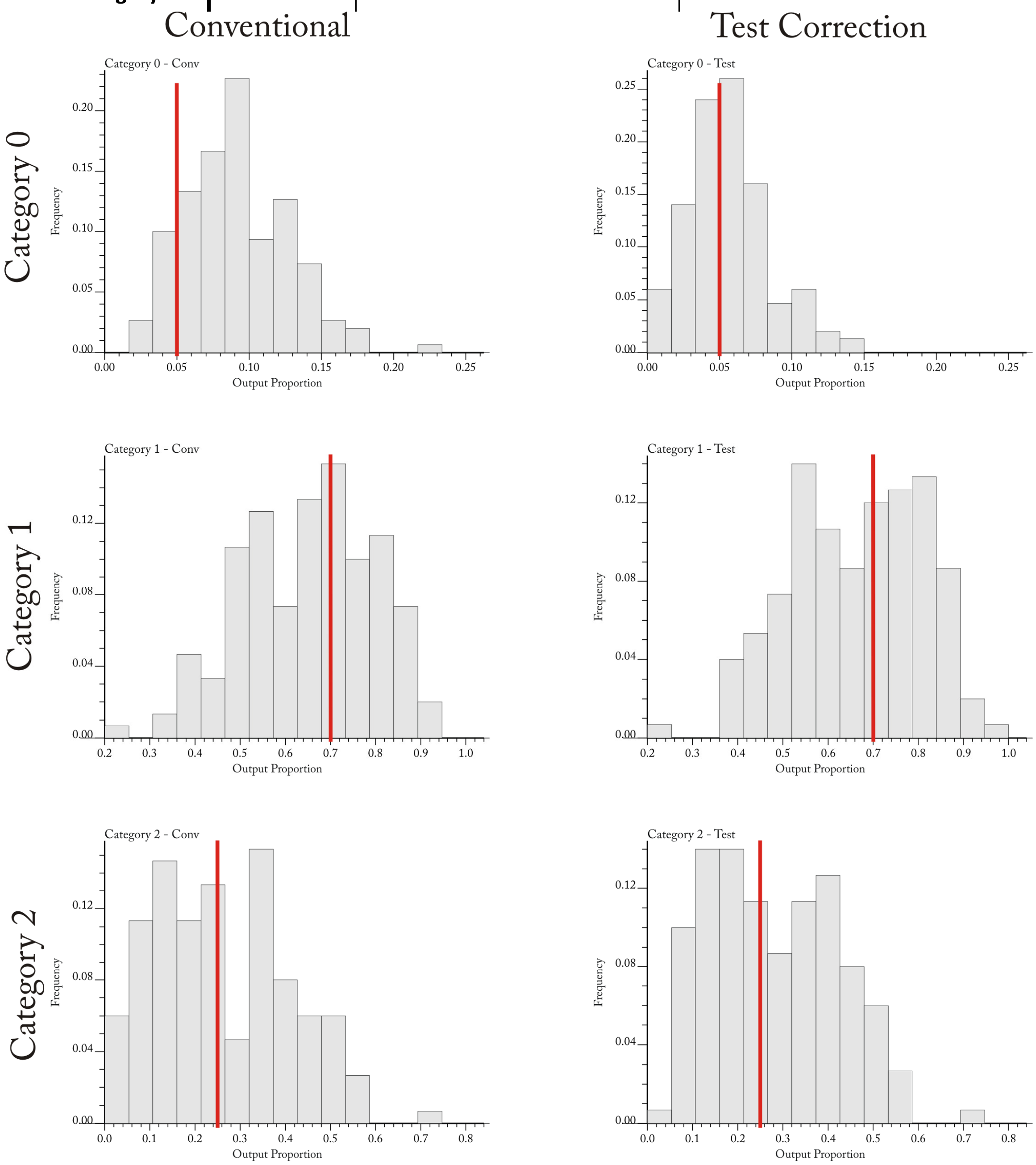

Figure 2: Histograms of output proportions for the conventional and test correction cases using *BlockSIS*

The most striking difference between the mean output proportions using the conventional correction and the test correction is for the first category. The conventional correction gives a mean value of 0.0918 which is almost twice as large as the input proportion of 0.05. The test correction gives a much more reasonable value of 0.0554. Other noticeable differences are that the test correction gives a lower standard deviation in output proportions. The test correction does not improve on the conventional correction for category 2, giving a mean of 0.2827 compared to 0.2574.

Discussion

For this case study, the proposed correction generally improves over the conventional correction with regards to reproducing an input proportion. All category means are improved upon with the exception of category 2. While the output proportion for category 2 is slightly worse using the proposed correction, the relative difference between 0.28 and 0.25 is much smaller than the relative difference between 0.09 and 0.05. The decrease in standard deviation is notable and is desirable, but worthy of further investigation.

The effect of order relation deviations and the potential for the introduction of a bias is significant. The proposed correction does not treat negative probabilities unfairly and generally produces better results for the case study conducted in this small study.

References

Deutsch, C.V., 2005, A Sequential Indicator Simulation Program for Categorical Variables with Point and Block Data: *BlockSIS*, *Centre for Computational Geostatistics,* Volume 7, 402.

Deutsch, J.L. and Deutsch, C.V., 2009, Checking and Correcting Categorical Variable Trend Models, *Centre for Computational Geostatistics*, Volume 11, 133

Journel, A.G., 2002, Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses, *Mathematical Geology*, Volume 34, 5, pp 573-596.