Paper 405, CCG Annual Report 11, 2009 (© 2009)

Debiasing with Multiple Soft Secondary Data

Sahyun Hong and Clayton V. Deutsch

Obtaining representative statistics for geostatistical modeling is important; however, preference in the sample statistics is unavoidable. In this work, debiasing the global proportion of categorical variables with soft secondary data is developed. The joint relation among primary and secondary variables is accounted for using nonparametric technique. Marginal condition of the modeled joint pdf is evaluated and debiased global proportions are derived from the modeled joint distribution. Programs implement the methodology and examples are shown.

Introduction

Geostatistical models must reproduce the input statistics and spatial continuity; however, there is no intrinsic declustering or debiasing techniques to obtain representative statistics in geostatistical modeling methods. Sample clustering occurs when the domain of interest is sampled preferentially. It is natural that spatial data is collected in a non-representative way. For example, high reservoir quality area has more exploration wells than the low quality area. If all of samples are combined with equal influence to determine the distribution, the high quality part will have too much influence.

Declustering is widely used to correct the biased statistics: global mean and variance. Various types of declustering methods such as cell, polygonal and kriging weight declustering have been developed (Isaaks and Srivastava, 1989). These declustering methods can correct the bias inherent in sample statistics caused by geometric clustering when there are sufficient data in all areas of the domain.

Debiasing uses prior knowledge or quantitative secondary data for correcting the biased statistics. Soft secondary data may be representative of entire area of interest and relation between primary and secondary variable can be modeled in a either parametric or nonparametric way. The central idea of debiasing is to extrapolate the primary distributions over full range of secondary values.

This work addresses the debiasing method with secondary variables to obtain a representative proportion of categorical variable. Debiasing with a single secondary variable is straightforward. Benefits of the work dwell in using several secondary data for correcting the bias and accounting for the non-Gaussian relations. The joint relation between primary and secondary variables is modeled using kernel density estimator. The modeled joint pdf is updated under the constrained marginal conditions. The debiased proportion is then derived from the integrating the updated joint pdf over the outcomes of secondary variables.

Methodology

Debiasing uses soft data that are representative of the entire area, and an understanding of the relationship between the primary and secondary data to correct the primary distribution. Let us denote the categorical variable indicating facies or rock type S taking one of s=1,…,K, and secondary variable $\mathbf{Y}=[y_1,\dots y_{\mathrm{nsec}}]$. The global proportion of categorical variable can be debiased with respect to the secondary data distribution such as:

$$p_{\mathrm{S}}^{debias}(s)=\int_{-\infty}^{\infty} f_{\mathrm{SY}}(s,\mathbf{y})d\mathbf{y} \tag{1}$$

where the $f_{\mathrm{SY}}(s,\mathbf{y})$ is a modeled joint distribution that satisfy all axioms of joint pdf: positive densities, $\sum_{\forall s}\int f_{\mathrm{SY}}(s,\mathbf{y})d\mathbf{y}$ = 1 and reproduction of lower order marginal distribution. Use of linear assumption among secondary variables is straightforward. Challenges are to use secondary data with accounting for non-Gaussian relations among themselves. Kernel density estimator is used for the modeling of joint relation among variable S and **Y**. Choice of kernel function types is less critical; however, kernel bandwidth has critical influence on the resulting density estimates. Scott (1992) suggested data driven kernel bandwidth and this work adopts the analytical suggestion as a guideline. The implemented program gives an option to change the kernel bandwidth based on the analytically suggested bandwidth.

Before applying the debiasing equation (1), one should evaluate some axioms of joint distribution: positiveness of estimated densities, closure condition $\sum_{\forall s}\int f_{\mathbf{SY}}(s,\mathbf{y})d\mathbf{y}$ and reproduction of lower order marginal distribution. Kernel density estimator meets the first two conditions. Marginal condition the modeled joint pdf must meet is following:

$$\sum_{s=1,\ldots,K} f_{\mathbf{SY}}(s,\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}) \tag{2}$$

which states that integration of joint distribution over possible outcomes of primary variable S should amount to the secondary data distribution $f_{\mathbf{Y}}(\mathbf{y})$. This condition is often violated because the modeling of $f_{\mathbf{SY}}(s,\mathbf{y})$ is performed with limited well samples, however, the modeling of $f_{\mathbf{Y}}(\mathbf{y})$ is done with the exhaustive samples. To meet this marginal condition, a simple scaling procedure is advanced:

$$\left(\frac{f_{\mathbf{Y}}(\mathbf{y})}{\sum_{s=1,\ldots,K} f_{\mathbf{SY}}^{(0)}(s,\mathbf{y})}\right)\times f_{\mathbf{SY}}^{(0)}(s,\mathbf{y}) = f_{\mathbf{SY}}^{(1)}(s,\mathbf{y}) \tag{3}$$

The ratio in the parenthesis becomes 1 if the marginal condition in (2) is met, otherwise the initial joint pdf $f^{(0)}$ is modified by the amount of the calculated ratio. Corrected joint pdf is denoted as $f^{(1)}$ to differentiate initial distribution. This scaling procedure directly accounts for differences from reference and reproduced marginal distribution.

In summary, the proposed debiasing approach is based on three steps: (1) build the joint distribution of all secondary variables, build the initial joint distribution of the primary and all of secondary variables, (2) correct the initial joint distributions under the secondary marginal distributions, and (3) add up the corrected joint densities over the range of secondary values:

$$p_{\mathrm{S}}^{debias}(s) = \int_{\mathbf{y}}\left(\frac{f_{\mathbf{Y}}(\mathbf{y})}{\sum\limits_{s=1,\ldots,K} f_{\mathrm{SY}}^{(0)}(s,\mathbf{y})}\times f_{\mathrm{SY}}^{(0)}(s,\mathbf{y})\right)d\mathbf{y} \tag{4}$$

Arithmetic operation in the parenthesis is for correcting the $f^{(0)}$ by secondary marginal distribution $f_{\mathbf{Y}}(\mathbf{y})$. Integrand operator outside the parenthesis adds up the corrected joint densities over possible outcomes of secondary values, leading to the debiased proportion $p_S(s)$.

Program Description

Two programs are implemented for the debiasing with multiple secondary data. The program secMDE.exe is for modeling the secondary data distribution in a nonparametric way. Kernel density estimator is implemented for the nonparametric modeling. An example parameter file is shown below:

```
line 1:  2           -Number of sec variables
line 2:  1  2        -column number of sec var
line 3:  sec.out     -secondary data file
line 4:  30          -number of bins
line 5:  0.8         -smoothing factor
line 6:  secpdf.out  -output secondary pdf file
```

Figure-1: An example of secMDE.exe parameter file

Line 1 – 3 defines number of secondary variable under consideration. Line 4 defines the number of bins where joint densities are modeled. The program uses the analytical suggestion as kernel bandwidths. One can adjust the smoothness of the modeled pdf in line 5; small value makes the modeled pdf less smooth and large value makes the modeled pdf more smooth. Output file is defined in line 6.

The second program (debcat.exe) is for modeling the joint distribution of primary and secondary variables, and updating the initial joint distribution with the modeled secondary data pdf.

```
line 1:  2           -number of sec variables
line 2:  3           -number of categories
line 3:  1 2 3       -code
line 4:  well.dat    -well data
line 5:  3 4 5       -column for primary and secondary var
line 6:  secpdf.out  -secondary data pdf file
line 7:  30          -number of bins
line 8:  0.8         -smoothing factor
```

Figure-2: An example of debcat.exe parameter file

The number of secondary variables, categories and codes are defined in the line 1 – 3. Well data including categorical primary and secondary data is specified in the line 4 and 5. The secondary data pdf file is defined in line 6. Notice that the number of bins defined in line 7 must be same as bin number specified in the secondary data pdf modeling (line 4 in the figure-3). The program writes out the debiased proportion of each categories on a screen.

Examples

Binary reference image is prepared for comparing debiased and true statistics. True global means are 0.453 and 0.547, respectively for code 0 and 1. A total of 16 well data are sampled from the reference image and samples are separated by 64 unit equal interval. In early stage of reservoir appraisal, few wells with little preferential drilling are often cases. This synthetic example demonstrates that case.

The naïve global means are 0.313(=5/16) and 0.688(=11/16) for each code. Naïve statistics are biased by more than 26% compared with true ones. Figure-3 shows reference image, sample data locations and simulated secondary data. Secondary data (Y) is simulated to well reflect the true facies distribution: low values tend to predict the code 0 and high values tend to predict code 1. Any declustering methods wholly depends on the geometric configuration of sample data locations would provide global mean almost same as naïve means. For example, declustered means are changed very little with respect to the change of cell sizes generating 0.325 and 0.675 as a declustered proportions which are still biased by 23% based on true values.

Debiasing with an exhaustive secondary data is applied. The secondary data distribution $f_Y(y)$ is first obtained using secMDE.exe program. A large enough bin number (50) is specified and analytical kernel bandwidth multiplied by 0.85 is used for the distribution modeling. Initial bivariate distribution (before correction with marginal condition) is modeled and shown in Figure-4. Due to sample data paucity, the modeled distribution curves appear smooth. The modeled $f_{SY}(s=0,y)$ and $f_{SY}(s=1,y)$ are shown as a smooth solid line. The distribution of collocated secondary data is represented as a bar chart.

Initial distributions are modified under the marginal constraint of the secondary data distribution $f_Y(y)$. Figure 5 shows the modified distributions that have detailed variations in distribution. Those distributions exactly amount to the secondary data distribution as shown in the bottom of the figure. Debiased global proportions of each code are obtained by (numerically) summing up the shaded area below the updated curves in Figure-5. Global proportions are summarized in the below for comparison:

	True	16 Samples with equal weights	Cell declustering	Debiasing
Code 0	0.453	0.313	0.325	0.415
Code 1	0.547	0.688	0.675	0.585

Cell declustering barely corrects the bias because declustering techniques inherently assume the sample data cover full range of data values. Declustering methods assume bias in samples arise only by spatial clustering, not by data value clustering. The demonstrated example; however, shows that collocated secondary data values are limited to [-1.99,0.41] and [-1.7,1.92] for code 0 and 1, respectively. The discussed debiasing method is based on the use of secondary data which are distributed over the full range of secondary values, say [-3.04,1.92] for this example. The bias caused by clustering within the limited secondary values is mitigated by the comprehensiveness of secondary data.

Another simulated secondary variable Y_2 is added to Y_1 for debiasing. Their correlation is about 0.66. Implemented programs secMDE.exe and debcat.exe were applied for using two secondary data. Parameter files are set appropriately. Debiased proportions are compared in the table below.

	True	16 Samples with equal weights	Cell declustering	Debiasing with Y1	Debiasing with (Y1+Y2)
Code 0	0.453	0.313	0.325	0.415	0.433
Code 1	0.547	0.688	0.675	0.585	0.567

Conclusions

Non-representative sampling is unavoidable in most geostatistical modeling cases. Exhaustive secondary data are often available and they are valuable information source to represent the entire area. The main focus of this work was to consider secondary data to obtain the debiased proportion of categorical variable. A non-parametric kernel density estimation method was applied to obtain the joint distribution among primary and secondary variables. Marginality condition is evaluated and a simple scaling procedure is advanced to meet the condition. Proportion of primary variable is then obtained from numerical summing up of the corrected joint distribution.

E. H. Isaaks, R. M. Srivastava, 1989, An introduction to applied geostatistics, New York, Oxford University Press.

D. W. Scott, 1992, Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley and Sons, Inc., New York.

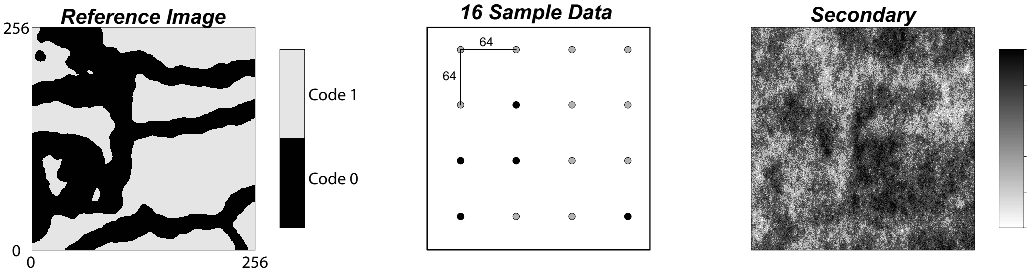

Figure 3: Prepared reference image, well sample locations extracted from reference image and a simulated secondary data. Sample means are biased by 26% or above based on the true statistics.

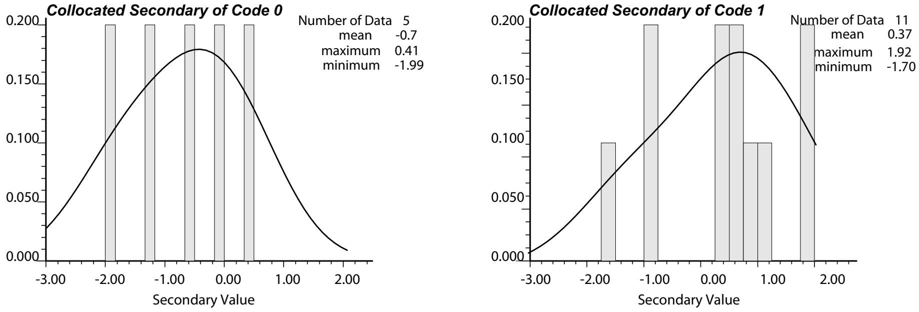

Figure 4: Experimental data distribution and the modeled distribution $f_{SY}(s,y)$ based on sample data. The smooth line is a kernel estimator of bar chart.

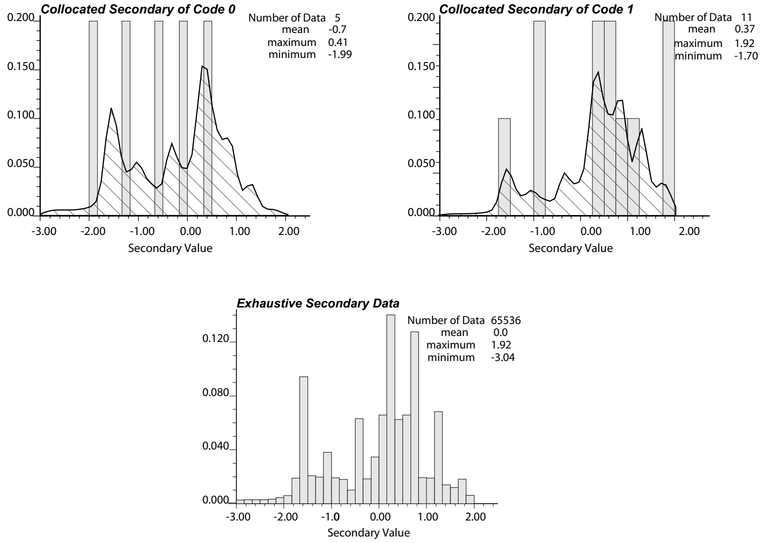

Figure 5: The corrected bivariate distributions by an imposed marginal condition (top). A marginal distribution of secondary data is shown in the bottom. Sum of solid lines shown in the top exactly amount to the secondary data distribution.