Generalized Linear Mixed Model with Spatial Covariates

by Alex Zolot (Zolotovitski)
StatVis Consulting

alex@zolot.us alexzol@microsoft.com
Introduction

• The task:
• Two Traits of subjects (plants) depends on
  1) Type (variable Entry_Name) and 
  2) Location in 2D Fields (Field, Row, Column).
• Dependence of Type – fixed effect, on Location – random effect.
• All locations are different, but similarity decrease with distance.
Parts of Solution:

- Descriptive statistics and visualization.
- Data preparation.
- Building the model.
- Validation.
- Programming.
- Automation, GUI
- Optimization of experimental design
Building the Model.
Type – Location Decomposition

• If the attribute value collected on an experimental unit (cell) is represented by the term $Y$, then the attribute can be generally modeled as follows:
  \[ Y = T + L + Err. \]

• In general liner model (GLM) $Y$ is linked to original variable Trait (Trait1 or Trait2) by linking function $g()$:
  \[ Y = g(Trait) \] (1)
  \[ Y = T + L + Err \] (2)
Box-Cox optimization

We looked for $g()$ in form of Box-Cox transformation that maximize average by Entry_Name p-value of test Shapiro for normality.

The result of this procedure

<table>
<thead>
<tr>
<th>Fun:</th>
<th>I</th>
<th>log(x)</th>
<th>$x^{1/3}$</th>
<th>$\sqrt{x}$</th>
<th>$x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro p.value:</td>
<td>0.37635</td>
<td><strong>0.52564</strong></td>
<td>0.49668</td>
<td>0.47207</td>
<td>0.17314</td>
</tr>
</tbody>
</table>

For simplicity we use $\lambda = 0$ corresponding to variable $Y = \log(\text{Trait})$ that has almost highest normality, but easier for understanding.
• Tests for homoscedasticity also confirmed advantage of logarithmic linking function in glm.
• So in our program we use log linking $Y = \log(Trait)$ with following variables names:

$$\text{Tra} = \text{Trait1 or Trait2}$$

$$\text{LTra} = Y = \log(Trait)$$

• with type – location decomposition

$$Y = Y_{ty} + Y_{loc} + \text{res}$$

$$\text{Tra} = \text{Tra}_{ty} \times \text{Tra}_{loc} + \text{noise}$$

• where

$$\text{Tra}_{ty} = \exp(Y_{ty}) \text{ and } \text{Tra}_{loc} = \exp(Y_{loc})$$

• In our case type “ty” is related to variable Entry_Name and location “loc” to tuple (Testing_Site, Field, Row, Column).
**Iteration of Type – Location decomposition.**

To get decomposition (2), we use the following iterative procedure:

\[ Y = Y(type, \text{loc}) = Y_0 = \log (\text{Trait}) \]

Do until convergence:

\[ Y_{old} = Y \]

\[ T(type) = \text{mean}(Y \mid \text{Type} = \text{type}), \] where Type = EntryName

\[ L_0 = Y - T(type) \]

For each TSF, using `krige.cv` package `gstat`:

\[ L(\text{loc}) = \text{cv.Predict(Krig(L0 ~ Row + Column, \text{loc}, \theta))} \]

\[ Y_{new} = Y_0 - L(\text{loc}) \]

\[ Y = (1 - \lambda) \cdot Y_{old} + \lambda \cdot Y_{new} \]

Loop until \[ \|Y_{new} - Y_{old}\| < \varepsilon \]

\[ T(type) = \text{mean}(Y \mid \text{Type} = \text{type}) \]

where \( \theta \) is the set of parameters of kriging that we have to optimize, and \( \lambda \) is parameter of acceleration.
• We control SSE (sum of squares of residuals) and after it differences becomes smaller than tolerance or after fixed number of “burn out” cycles we get mean and standard deviation of $Y_{loc}$ and $Y_{ty}$:

$$Y_{loc}.m = \text{mean}(Y_{loc} \mid \text{burnOut} < \text{iter} \leq \text{maxiter})$$

$$Y_{loc}.sd = \text{sd}(Y_{loc} \mid \text{burnOut} < \text{iter} \leq \text{maxiter})$$

• Residuals depend on Row, Column after excluding Type and Test_Site components:

```r
library(nlme)
fm1 <- lme(LTra ~ Entry_Name, sds, random = ~ 0 | Entry_Name)
  # effect of Testing_Site ======
  sds$resid1 = fm1$resid[,1]  # now means by Entry_Name are excluded
fm2 <- lme(resid1 ~ TSF, sds, random = ~ 0 | TSF)  # not necessary, just to exclude mean by TSF.
  sds$resid2 = fm2$resid[,1]  # now means by TSF are excluded
```
Fig.2. Excluding Type-dependence in 0-approximation.
Kriging cross-validation and optimization.

- Two kriging parameters – range and nugget
- Methods of Nelder and Mead (1965)
- Optimization of kriging parameters is very important and time-consuming procedure, so our results must be considered as preliminary.
- Linear regression on residuals with predictors Row and Column, that we considered as numerical variables – so all our prediction on this stage used only 4 kriging adjustment parameters – sill, range, nugget, and anisotropy.
We also tried to use regression with Row and Column as random effects, but found that additional degrees of freedom increase AIC:

```r
ds$cRow=paste('r',ds$Row, sep='')
ds$cCol=paste('c',ds$Column, sep='')

lm00= glm( resid2 ~ var1.pred, data = ds)
lm0= glm( resid2 ~ var1.pred + Column + Row , data = ds)
lmR= glm( resid2 ~ var1.pred + Column + Row + cRow , data = ds)
lmC= glm( resid2 ~ var1.pred + Column + Row + cCol , data = ds)
lmRC= glm( resid2 ~ var1.pred + Column + Row + cCol+ cRow , data = ds)

c(AIC(lm00), AIC(lm0), AIC(lmC), AIC(lmR), AIC(lmRC))
# -3615.188 -3611.492 -3584.912 -3584.497 -3568.149
```
• Kriging on residuals after excluding Type effect in 0-approximation:

Fig. 4. Result of kriging + glm for TSF = 7231_F on Loc – dependant part of data
Variograms and anisotropy

Fig 6. Variograms for different angles for TSF = 7605_F5.
From Fig. 7 we see that elliptical model

\[ \text{variogram} \left( \text{diffRow}, \text{diffColumn} \right) = f \left( \frac{(\text{diffRow})^2}{a^2} + \frac{(\text{diffColumn})^2}{b^2} \right) \]

with one parameter of anisotropy

\[ \text{anis} = \frac{b}{a} \]

is not very good fitting for anisotropy but in standard kriging procedures only this model of anisotropy is implemented. To improve accuracy of our model in future we could use a multistep approach to overcome this inaccuracy of elliptical model.
Choosing number of iterations.

Fig. 10. $\ln(\text{SSE})$ vs iteration for different acceleration parameter $\lambda_a = \lambda$. 
As a result sharpness of signal increased essentially:

Fig.9. Density for distribution Tra and Tra_Ty for Trait=1, Treatment =2.
Programming. Automation, GUI

- R for analyzing and modeling
- packages 'stats', 'sqldf', 'spatstat', 'gstat', 'sp', 'lattice', 'tcltk', 'tkrplot', 'graphics'

Fig.15. Screenshot of GUI
Conclusion

• Noise: The sum of the squared residuals of the model should be minimized.

Resulting SSE:

<table>
<thead>
<tr>
<th>Dataset 1: Treatment</th>
<th>Trait</th>
<th>SSE</th>
<th>SST</th>
<th>Rsq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>14.308</td>
<td>146.087</td>
<td>0.902</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>48.392</td>
<td>191.456</td>
<td>0.747</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset 2: Treatment</th>
<th>Trait</th>
<th>SSE</th>
<th>SST</th>
<th>Rsq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>40.769</td>
<td>286.499</td>
<td>0.858</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>80.945</td>
<td>317.27</td>
<td>0.745</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>35.998</td>
<td>150.875</td>
<td>0.761</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>58.341</td>
<td>175.262</td>
<td>0.667</td>
</tr>
</tbody>
</table>
• Parsimony: Fitted parameters for location-based artifacts must comprise a relatively small portion of the total number of parameters.
  
  We used only 4 fitting parameters of kriging for each (Treatment, Trait, TSField)

• Signal: The remaining signal in the dataset should be maximized, as measured by a statistical test to differentiate the entries.
  
  Sharpness of signal increased essentially, as Fig.8-9 shows.

• Dropped values: The amount of dropped data values should be kept to a minimum.
  
  We dropped about 1% as outliers.

• Speed and ease of use: Some automation with an intuitive user-interactive interface.
  
  Our GUI has only 6 buttons in Tcl/Tk and only one button in RExcel.
Next Steps

• The performance could be improved essentially if we combine iterations with cross-validation. Results that we delivered were obtained with 20 fold cross-validation and 19 iterations, that means dataset was scanned $20 \times 19 = 380$ times and it took about 154 min. If we combine iterations with cross-validation, we estimate to reach the same accuracy in about 40 scans, that is 10 time faster, so it would take less than 1 min.

• We can also improve accuracy by using two stage kriging to extend managing of anisotropy in our variogram model from 1-parameter ellipse with main axis in column direction to at least 2-parameters of two ellipses in column and row direction or in arbitrary angle. We estimate possible accuracy improvement in about 25-30% decrease of SSE.
Generalized Linear Mixed Model with Spatial Covariates

by Alex Zolot (Zolotovitski)

alex@zolot.us  www.zolot.us