Definition of Region for Clinical Trials

Mathematical Optimization Approach

Alex Zolot (Zolotovitski), PhD

Alex.Zolot@StatVis.com
alexzol@microsoft.com

Yoko Tanaka
Eli Lilly and Company
Alex Zolot (Zolotovitski), PhD

Senior Researcher at Microsoft (Bing).

Formerly - a Senior Statistician / Sr. Engineer at Sun Microsystems, an Analytic Science Manager at FICO, executed successful data mining projects for Kraft Foods, Visa, Discover Financial, Cox Communications.

Fields of expertise include statistical analysis and modeling, and data mining, mainly predictive analytics.

Ph.D in Theoretical and Mathematical Physics, and a Ph.D. in Economics.

Certified Advanced SAS Programmer, MCP.

www.zolot.us
• **InnoCentive** - [www.Innocentive.com](http://www.Innocentive.com) - provides connection services between "Seekers" and "Solvers." Seekers are the companies searching for solutions to critical challenges. Solvers are the 185,000 registered members of the InnoCentive crowd who volunteer their solutions to the Seekers. Solvers whose solutions are selected by the Seekers are compensated for their ideas by InnoCentive, which acts as broker of the process.

• **Eli Lilly** posted the challenge, “to identify proposals for new regions for clinical trials which are supported by information which is currently publicly available (publication, clinicaltrials.gov, medline, etc.).
• The task of definition of regions is universal for different actions – Clinical Trials or Soccer Cup.

• No universal solutions for clinical trials (CTReg)
  
  Optimal regions depends on many parameters, so in our solution we describe not definition of regions, but *procedure* how to get the definition if we have required parameters.

• We also attach small program in R that generates regions given a set of parameters (‘weights”) provided by experts.
Stage 1 – Experimental Design (DoE)

Typically clinical trials may be designed to do assess the safety and effectiveness of some medications or devices on a specific kind of patient and could be formalized as analysis of response

\[ Y_i = f( p_j, x_k ) \]

health criteria \( Y \) on treatment variables \( x \). The response depend on parameters \( p \) of patients, including mentioned by Seeker: culture, ethnicity, language, medical practice, patient/disease characteristic, regulatory filing system demographic and so on. The objective of clinical trials is reveal (find a good approximation) for function \( f \), that is a typical task of regression that could be non-linear and non-parametric. The regression procedure is out of scope of this work. For simplicity let us suppose linear regression

\[ Y_i = \sum a(p_j) \times x_k \]

then objective of clinical trials is to estimate \( a(p_j) \). For definition of CTRegs is convenient to aggregate parameters \( p_j \) into one aggregated variable \( p \) (“aggregated parameter”, AP).

**Result of DoE** – sample/block size - supposed to be done – out of the scope.
UN Regions – 5 regions/23 subregions, ~ 200 nodes (countries)

Why 5 or 23 regions? Optimal Number of Regions -?
Optimal Number of Regions -?

Too many?  

Too small?
Optimal Number of Regions $n$ -?

- Minimization of

$$\text{Cost} = \text{AdminCost} (n_+) + \text{SizeCost} (R_+) + \text{NonUnifCost} (R_+)$$

“+” indicates increasing function,

$R$ is “average” size of region
Optimal Number of Regions \( n \) -?

\[
\text{Cost} = \sum_{u \in CTR} (A_u + B_u n_u + C_u n_u < (R - R_{cu})^2 > ) \tag{2}
\]

\[
= \sum_{u \in CTR} (A_u + B_u n_u + C_u n_u \text{Var}(R)_u) \tag{3}
\]

where \( \Sigma \) means summation by Regions \( u \),
\( n_u \) – number of nodes in Region \( u \),
\( A_u, B_u, C_u, D_u \) are cost parameters that depends on Regions \( u \) and
\( R_u \) are known constants - average size of Region \( u \) that could be estimated as \( \text{AREA}_u^{1/2} \).
Optimal Number of Regions \( n \) -?

Minimization of

\[
\text{Cost} = \sum_{u \in \text{CTR}} (A_u + B_u n_u + C_u n_u \text{Var}(R)_u) \quad (4)
\]

\[
\text{if coefficients do not depend on regions}
\]

\[
= A n_{\text{reg}} + B n_{\text{nodes}} + C n_{\text{nodes}} \text{mean(Var(R))} \quad (5)
\]
R (geo) $\rightarrow$ General set of parameters $\Sigma w_j p_j$

Cost $= A n_{reg} + n_{countries} \text{mean}(\Sigma w_j^2 \text{Var}(p_j))$ (6)

$= A n_{reg} + \Sigma w_j^2 (p_j - p_{cj})^2$ (7)

$A$ – cost to create one Reg.Center

$w_j^2$ – cost of intra-region variance of $p_j$

$p$ - raw variables or
- Indexes of raw variables or
- Factors build on raw variables

Categorical variables $\rightarrow$ dummy numerical variables
Minimization = Clustering

\[ \text{Min}(\text{Cost}) = \min(A n_{\text{reg}} + \sum w_j^2 (p_j - p_{c_j})^2) \quad (8) \]

= typical task of clustering
Minimization = Clustering

$$\text{Min}(\text{Cost}) = \min( A n_{\text{reg}} + \sum w_j^2 (p_j - p_{cj})^2 )$$  \hspace{1cm} (8)

= typical task of clustering in data mining

We can start from wide set (hundreds or thousands) parameters, characterizing countries (nodes). Then dimension reduction - e.g. via factor (PCA) analysis.

In our simple mock-up we use GDP, Human Development Index, Mortality Rate of Cancer and Tuberculosis Treatment Success.
14 CTRregs, equal weights, Weighted Average
14 CTRregs, HDI only, Clusters (k-means),
14 CTRegs, HDI only, Clusters (pma), labels on
No geo, 3 Categorical variables -> 7 dummy, 1st factor, 10 levels
Output for PCA Analysis: Correlation Circles.
Output: Table of CTRregs.

<table>
<thead>
<tr>
<th>Country</th>
<th>CTReg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grenada</td>
<td>5</td>
</tr>
<tr>
<td>Greenland</td>
<td>NA</td>
</tr>
<tr>
<td>Germany</td>
<td>3</td>
</tr>
<tr>
<td>Guam</td>
<td>NA</td>
</tr>
<tr>
<td>Greece</td>
<td>4</td>
</tr>
<tr>
<td>Guatemala</td>
<td>8</td>
</tr>
<tr>
<td>Guinea</td>
<td>8</td>
</tr>
<tr>
<td>Guyana</td>
<td>7</td>
</tr>
<tr>
<td>Haiti</td>
<td>9</td>
</tr>
<tr>
<td>Honduras</td>
<td>7</td>
</tr>
<tr>
<td>Croatia</td>
<td>6</td>
</tr>
<tr>
<td>Hungary</td>
<td>6</td>
</tr>
<tr>
<td>Iceland</td>
<td>2</td>
</tr>
<tr>
<td>India</td>
<td>8</td>
</tr>
<tr>
<td>Iran (Islamic Republic of)</td>
<td>2</td>
</tr>
<tr>
<td>Israel</td>
<td>0</td>
</tr>
<tr>
<td>Italy</td>
<td>3</td>
</tr>
<tr>
<td>Cote d'Ivoire</td>
<td>6</td>
</tr>
<tr>
<td>Iraq</td>
<td>9</td>
</tr>
<tr>
<td>Japan</td>
<td>0</td>
</tr>
<tr>
<td>Jamaica</td>
<td>6</td>
</tr>
<tr>
<td>Jordan</td>
<td>3</td>
</tr>
<tr>
<td>Kenya</td>
<td>4</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>5</td>
</tr>
</tbody>
</table>
• Why we need PCA?

• Generalization: Hierarchy – hierarchical clustering.
Conclusion

We developed

- General optimization approach to create optimized CT Regions using standard clustering technics and expert defined weights (or cost coefficients) for arbitrary set of important imput variables

- Mock-up R application implementing the technic and visualizing results of clustering, levels of factors (or main principal components) or raw variables
Thank you