Making the Martyn Method a Reality

Judy Brown, Ray Martyn, Forrest Nutter, Jim Stack & Carla Thomas

Neil McRoberts
Plant Pathology Department
UC Davis
nmcroberts@ucdavis.edu
Aims

• Two inter-connected aims:
  • (A) Examine whether there are generic types of disease
  • (B) Examine the extent to which well-informed raters agree on the important properties of diseases

  – Note, A is an emergent property of the extent to which B is true.
Algorithm design

• A blend of human and computer components

Double-blind approach: analyst did not know identities of raters or diseases in the first iteration
Data structure

4 Raters × 14 Diseases × 75 variables (binary - 1/0)

$2^{75} \approx 10^{21} – 10^{22}$ possible unique fingerprints
Individual rater disease dendrograms

Farthest neighbour

Simple Matching Coefficient

1 (LWA)
5 (LWRB)
3 (CVC)
13 (SPBR)
6 (HLB)
8 (PPV)
2 (LWC)
11 (Rathyibac)
4 (DMC)
10 (RedBl)
9 (PWart)
14 (WSR)
7 (Pkern)
12 (Rsr3b2)
Alternative geometric interpretation

Can think of 3 variables, say, as defining $2^3 = 8$ unique locations in a 3D space, or as the 8 vertices on a cube with sides of length = 1.

We have a $2^{75}$ hypercube.

The Martyn theorem is that disease will occupy a limited set of close vertices and raters will put diseases on similar sets of vertices.
Projection of “distances” between diseases according to four rates
Generalized Procrustes analysis

Procrustes analysis is a technique for matching two sets of points to each another and finding which points are responsible for lack of fit. Generalized Procrustes analysis extends the approach to more than two data sets.

Can calculate point by point distance once the best –possible overall fit has been found.

Distances (or residuals) calculated as “squared” values so sum of squares gives overall lack of fit and each point can be attributed a %S.S.
Considering all raters together
ANOVA associated with GPA

<table>
<thead>
<tr>
<th></th>
<th>fit</th>
<th>residual</th>
<th>total</th>
<th>%residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWA</td>
<td>5.817</td>
<td>2.578</td>
<td>8.396</td>
<td>10.844</td>
</tr>
<tr>
<td>LWC</td>
<td>4.134</td>
<td>0.274</td>
<td>4.409</td>
<td>1.154</td>
</tr>
<tr>
<td>CVC</td>
<td>8.957</td>
<td>0.554</td>
<td>9.511</td>
<td>2.330</td>
</tr>
<tr>
<td>DMC</td>
<td>8.415</td>
<td>1.782</td>
<td>10.198</td>
<td>7.497</td>
</tr>
<tr>
<td>LWRB</td>
<td>8.547</td>
<td>3.476</td>
<td>12.023</td>
<td>14.621</td>
</tr>
<tr>
<td>HLB</td>
<td>9.173</td>
<td>1.222</td>
<td>10.395</td>
<td>5.141</td>
</tr>
<tr>
<td>Pkern</td>
<td>0.029</td>
<td>2.989</td>
<td>3.018</td>
<td>12.573</td>
</tr>
<tr>
<td>PPV</td>
<td>6.292</td>
<td>0.415</td>
<td>6.707</td>
<td>1.746</td>
</tr>
<tr>
<td>PWart</td>
<td>3.945</td>
<td>1.072</td>
<td>5.017</td>
<td>4.510</td>
</tr>
<tr>
<td>RedBl</td>
<td>6.380</td>
<td>1.605</td>
<td>7.985</td>
<td>6.749</td>
</tr>
<tr>
<td>Rathayibac</td>
<td>3.232</td>
<td>2.477</td>
<td>5.709</td>
<td>10.417</td>
</tr>
<tr>
<td>Rsr3b2</td>
<td>3.586</td>
<td>2.282</td>
<td>5.868</td>
<td>9.596</td>
</tr>
<tr>
<td>SPBR</td>
<td>3.530</td>
<td>1.032</td>
<td>4.562</td>
<td>4.340</td>
</tr>
<tr>
<td>WSR</td>
<td>4.185</td>
<td>2.017</td>
<td>6.202</td>
<td>8.482</td>
</tr>
<tr>
<td>Totals</td>
<td>76.223</td>
<td>23.777</td>
<td>100.000</td>
<td></td>
</tr>
</tbody>
</table>

% fit S.S.

<table>
<thead>
<tr>
<th>rater</th>
<th>% fit S.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>21.39864</td>
</tr>
<tr>
<td>b</td>
<td>24.88994</td>
</tr>
<tr>
<td>c</td>
<td>28.09269</td>
</tr>
<tr>
<td>d</td>
<td>25.61872</td>
</tr>
</tbody>
</table>

Algorithm design

• A blend of human and computer components

Double-blind approach: analyst did not know identities of raters or diseases in the first iteration
Second iteration

• Raters work independently to reduce number of variables to more manageable, core set
• Reduced data sets compared with each other and original analysis to determine
  – If sufficient information has been retained
  – If consensus still exists
Second round results

- Reduced set of 36 candidate key items suggested by combining the lists from 4 raters.
- The level of agreement among raters was greater than would be expected by independent, random selection processes.
- Using the reduced data set to analyze similarity among the 14 diseases resulted in strong clustering of the 4 replicates of each disease. Evidence for the existence of generic disease types was retained in the reduced data set.
- The agreement between the whole- and reduced data disease similarities was very high (generally in the range of 95-100% fit S.S. in the GPA.)
Similarity analysis based on reduced data set
Second round compared with first
Item analysis

List D: Items to be discarded

1. Method of dissemination: Circulative in vector
2. Method of dissemination: Semi-persistent (in vector)
3. Over-seasons in soil saprophytically
4. Vector transmission: Circulative (passes through gut)
5. Exclusion: surveillance and monitoring
6. Protection: Not known if efficacious protective treatment exists
7. Protection: Efficacious treatment known but not feasible or practical
8. Resistance: begin a race-specific breeding program
9. Crop insurance available? (no)
10. Would public assistance/compensation to growers be available (no)
11. Would emergency declaration be a likely action (no)
12. Potential private economic impact (low)
13. Potential public economic impact (low)
14. Potential social impact (low)
Item analysis

List K3: Items with 3 votes to keep

1. Dissemination: fomite (human agency)
2. Overseasoning/survival: Null
3. Overseasoning: soil
4. Overseasoning: in vector
5. Vector transmission: trans-ovarial
6. Pathogen reproductive potential: Low/medium
7. Pathogen reproductive potential: medium/high
8. Host range: narrow - single species or genus
9. Host range: moderate - multiple species or genera within family
10. Geospatial pattern of host: contiguous
11. Area at risk: orchard/vineyard
12. Area at risk: natural landscape
13. Area at risk: forest
14. Area at risk: residential
15. Area at risk: greenhouse production
16. Latency period: short
17. Latency period: Long
18. Exclusion: testing/certification
19. Exclusion: sanitation
20. Exclusion: vector management
21. Eradication: host destruction
22. Eradication: Fumigation
23. Eradication: seed/plant part treatment
24. Resistance: resistance exists
25. Avoidance: is avoidance practical (yes)
26. Is there a probable trade impact (yes)
27. Is there a probable food safety impact (yes)
List K4: Items with maximum votes for retention

1. Dissemination: Wind
2. Dissemination: Rain
3. Dissemination: Seed/plant parts
4. Dissemination: Vectored
5. Dissemination: Soil
6. Disease cycle: polycyclic
7. Disease cycle: monocyclic
8. Host range: broad - multiple plant families
9. Area at risk: agricultural field
10. Exclusion: quarantine
11. Protection: known efficacious treatment exists
Summary

• The approach appears to have promise
• Four raters generated reasonably robust clustering of disease recovery plans
• We were able to identify a small number of variables that differentiate among disease types
• Prototype generic plans might be made from these variables and a set of others which all plans will include.