Introduction to SaTScan, seasonality and time series analysis

Hu Suk Lee

GIS training course for animal health workers
Jakarta, Indonesia
28 November 2016
Outline

• Introduction of SaTScan
• Examples
  • Leptospirosis in the USA
  • FMD in Korea

• Introduction of seasonality analysis
• Example
  • Viral Encephalitis in Vietnam

• Introduction of time series analysis
• Example
  • Brucellosis in South Korea
What is SaTscan?

• SaTScan is a freely available software that uses the scan statistic to detect clusters (www.satscan.org)

• To test whether a disease is randomly distributed over space, over time or over space and time.

• To perform geographical surveillance of disease, to detect areas of significantly high or low rates.
What is SaTscan?

Space
For each distinct window, calculate the likelihood, proportional to:

\[
\left( \frac{n}{\mu} \right)^n \left( \frac{N-n}{N-\mu} \right)^{N-n}
\]

\( n = \text{number of cases inside circle} \)
\( N = \text{total number of cases} \)
\( \mu = \text{expected number of cases inside circle} \)
Scan Statistics

• Circles of different sizes (from zero up to 50 % of the population size)

• For each circle a likelihood ratio statistic is computed based on the number of observed and expected cases within and outside the circle and compared with the likelihood $L_0$ under the null hypothesis.
• The **scan statistic** is the maximum likelihood over all possible circles
  – Identifies the most unusual clusters

• To find p-value, use Monte Carlo hypothesis testing
  – Redistribute cases randomly and recalculate the scan statistic many times
  – Proportion of scan statistics from the Monte Carlo replicates which are greater than or equal to the scan statistic for the true cluster is the p-value
What SaTScan can/can’t do?

**CAN**
- Identify spatial, temporal, spatial-temporal clusters
- Provide flexible geographic units

**CANNOT**
- Display maps of events and clusters locations
- Need GIS or mapping software (such as ArcGIS)
- Create other statistical and regression models
Introduction of Statistical models in SaTScan
Bernoulli Model

• There are animals with or without a disease (represented by a 0/1 variable)
• A set of cases and controls

• Purely temporal/spatial or the space-time scan statistics
Discrete Poisson Model

- The number of cases in each location is Poisson-distributed.

- Under the null hypothesis, and when there are no covariates, the expected number of cases in each area is proportional to its population size.

- Purely temporal, purely spatial and space-time

- This model a very good approximation to the Bernoulli model if few cases VS controls (less than 10%)
Space-Time Permutation Model

• Requires only cases data with information about the spatial location and time for each case (No information needed for population at risk)

• If the population increase (or decrease) is the same across the study region, that is okay, and will not lead to biased results

• The user is advised to be very careful when using this method for data spanning several years
  • population in some areas grows faster than in others
## Screenshot of SaTScan

### Input Section

<table>
<thead>
<tr>
<th>Case File:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control File: (Bernoulli Model)</td>
</tr>
<tr>
<td>Study Period:</td>
</tr>
<tr>
<td>Start Date: 2000 1 1</td>
</tr>
<tr>
<td>Population File: (Poisson Model)</td>
</tr>
<tr>
<td>Coordinates File:</td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>42.062</td>
<td>-72.6255</td>
</tr>
<tr>
<td>1002</td>
<td>42.3818</td>
<td>-72.4542</td>
</tr>
<tr>
<td>1007</td>
<td>42.2788</td>
<td>-72.4028</td>
</tr>
<tr>
<td>1008</td>
<td>42.1881</td>
<td>-72.5676</td>
</tr>
<tr>
<td>1020</td>
<td>42.1759</td>
<td>-72.5676</td>
</tr>
<tr>
<td>1028</td>
<td>42.0593</td>
<td>-72.4973</td>
</tr>
<tr>
<td>1033</td>
<td>42.2579</td>
<td>-72.5066</td>
</tr>
<tr>
<td>1035</td>
<td>42.3547</td>
<td>-72.5724</td>
</tr>
<tr>
<td>1036</td>
<td>42.0669</td>
<td>-72.4127</td>
</tr>
<tr>
<td>1039</td>
<td>42.4043</td>
<td>-72.6883</td>
</tr>
<tr>
<td>1040</td>
<td>42.2133</td>
<td>-72.6408</td>
</tr>
<tr>
<td>1057</td>
<td>42.0895</td>
<td>-72.3216</td>
</tr>
<tr>
<td>1060</td>
<td>42.3161</td>
<td>-72.6276</td>
</tr>
<tr>
<td>1069</td>
<td>42.1849</td>
<td>-72.3098</td>
</tr>
<tr>
<td>1077</td>
<td>42.0508</td>
<td>-72.7769</td>
</tr>
<tr>
<td>1085</td>
<td>42.1511</td>
<td>-72.7653</td>
</tr>
<tr>
<td>1095</td>
<td>42.1297</td>
<td>-72.4284</td>
</tr>
<tr>
<td>1096</td>
<td>42.4204</td>
<td>-72.776</td>
</tr>
<tr>
<td>1104</td>
<td>42.1351</td>
<td>-72.5702</td>
</tr>
<tr>
<td>1118</td>
<td>42.0951</td>
<td>-72.5259</td>
</tr>
<tr>
<td>1119</td>
<td>42.1239</td>
<td>-72.511</td>
</tr>
<tr>
<td>1201</td>
<td>42.4531</td>
<td>-73.2856</td>
</tr>
<tr>
<td>1230</td>
<td>42.1906</td>
<td>-73.406</td>
</tr>
<tr>
<td>1238</td>
<td>42.2951</td>
<td>-73.2268</td>
</tr>
<tr>
<td>1247</td>
<td>42.703</td>
<td>-73.0702</td>
</tr>
<tr>
<td>1301</td>
<td>42.628</td>
<td>-72.6009</td>
</tr>
<tr>
<td>1351</td>
<td>42.5363</td>
<td>-72.5154</td>
</tr>
<tr>
<td>1370</td>
<td>42.5933</td>
<td>-72.727</td>
</tr>
</tbody>
</table>
Input – Advanced Analysis

- Maximum Spatial Cluster Size:
  - 50.0 percent of the population at risk (≤ 50%, default = 50%)
  - 50.0 percent of the population defined in the max circle size file (≤ 50%)

- Spatial Window Shape:
  - Circular
  - Elliptic

- Use Isotonic Spatial Scan Statistic
Output

Text Output Format:
- Main Results File:

Geographical Output Format:
- KML file for Google Earth
- Shapefile for GIS software

Column Output Format:
- Cluster Information
- Stratified Cluster Information
- Location Information
- Risk Estimates for Each Location
- Simulated Log Likelihood Ratios/Test Statistics

ASCII | dBase
-----|-----
|     |     
|     |     
|     |     
|     |     
|     |     

Advanced >>
Executing
Example: Lepto
Leptospirosis

- A bacterial zoonotic disease caused by spirochetes of the genus *Leptospira*

- Pathogenic:
  - *L. interrogans* and *L. kirschneri*
    - More than 200 serovars
• Zoonotic disease
  – Transmitted to humans from a variety of wild and domesticated animal hosts
  – Most common reservoirs: rodents (rats), wild animals and farm animals in the US
  – Occupations that involve animal handling are more likely to contract disease

• Transmitted through damaged skin or mucus membranes of exposed humans and animals

• Indirect contact (water, soil and feed) with infected urine from an animal with leptospiruria
# Host animals

<table>
<thead>
<tr>
<th>Species</th>
<th>Common infections</th>
<th>Possible others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dogs</td>
<td>Canicola, Icterohemorrhagiae, Grippotyphosa, Pomona</td>
<td>Bratislava, Autumnalis</td>
</tr>
<tr>
<td>Cats</td>
<td>rarely identified</td>
<td></td>
</tr>
<tr>
<td>Cattle (and deer)</td>
<td>Hardjobovis, Pomona, Grippotyphosa, Icterohemorrhagiae</td>
<td>Australis, Autumnalis, Canicola, Bataviae, Hebdomadis, Krematosis, Tarassovi, Sejroe, Bratislava</td>
</tr>
<tr>
<td>Pigs</td>
<td>Pomona, Bratislava, Canicola, Tarassovi, Icterohemorrhagiae</td>
<td>Grippotyphosa, Sejroe</td>
</tr>
<tr>
<td>Sheep</td>
<td>Pomona, Grippotyphosa, Bratislava, Hardjo</td>
<td></td>
</tr>
<tr>
<td>Horses</td>
<td>Pomona, Bratislava, Canicola, Icterohemorrhagiae</td>
<td></td>
</tr>
</tbody>
</table>

Objective of study

- To evaluate spatial clustering by each serovar in the United States between 2008 and 2010 using data from IDEXX Laboratories Inc.
Materials and methods

• Procedures
  – Spatial clustering analysis (2008-2010)
    • Overall (any titers ≥ 1600) / each serovar
    • Spatial window was set at maximum 20%
    • $P$-value < 0.05

  – Geocoding was conducted from hospital zipcodes into geographical coordinates (latitude and longitude)
    • Sample locations with clusters were visualized on the map
Which model is appropriate?

- *Discrete Poisson Model*
- *May consider… Bernoulli Model*

Software programs

- Microsoft Excel & Note
- STATA version
- ArcGIS (ESRI, CA, USA)
- SaTScan
Results from Poisson and Bernoulli Models

- 50% of the population at risk area (serovar Autumnalis)
- Only primary cluster is significant in both models
Cluster map from Poisson Model

Submitted samples for canine leptospirosis between 2008 and 2010 (Any titers ≥1:1600)

- Sero-positive samples (Any titers ≥1:1600)
- All submitted samples
- Primary cluster
- Secondary cluster
Cluster map from Poisson Model

Submitted samples for canine leptospirosis between 2008 and 2010 (titers ≥1:1600)
Serovars Autumnalis, Grippotyphosa and Pomona

- Sero-positive samples (Any titers ≥1:1600)
- All submitted samples
- Primary cluster Autumnalis
- Primary cluster Grippotyphosa
- Primary cluster Pomona

0 500 1,000 2,000 Kilometers
Discrete Poisson Model

Submitted samples for canine leptospirosis between 2008 and 2010 (titers ≥1:1600)
Serovars Bratislava, Icterohaemorrhagiae and Canicola

- Sero-positive samples (Any titers ≥1:1600)
- All submitted samples
- Primary cluster Bratislava
- Secondary cluster Bratislava
- Primary cluster Icterohaemorrhagiae
- Primary cluster Canicola

Primary cluster for serovar Canicola
Secondary cluster for serovar Bratislava

Legend:
- Yellow circles: Primary cluster for serovar Canicola
- Red circles: Secondary cluster for serovar Bratislava
- Blue circles: Primary cluster Bratislava
- Purple circles: Secondary cluster Bratislava
- Small blue dots: All submitted samples
- Red dots: Sero-positive samples (Any titers ≥1:1600)
Cluster map from Poisson Model

Submitted samples for canine leptospirosis between 2008 and 2010 (titer ≥ 1600)
Serovars Bratislava, Icterohaemorrhagiae and Canicola

- Sero-positive samples (Any titers ≥ 1600)
- All submitted samples
- Primary cluster Bratislava
- Secondary cluster Bratislava
- Primary cluster Icterohaemorrhagiae
- Primary cluster Canicola
Spatial clusters of *Leptospira* seropositivity (MAT titers ≥1:1,600) between 2008 and 2010 in dogs in the United States (spatial window: max 20%)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Geographical area</th>
<th>Population</th>
<th>Radius (km)</th>
<th>Obs/exp = ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any titers ≥ 1600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,717; cases:1,487</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>Central Texas</td>
<td>2,139</td>
<td>232.87</td>
<td>262 / 162.02 = 1.62</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Secondary cluster</td>
<td>North-central Colorado</td>
<td>83</td>
<td>7.94</td>
<td>22 / 6.29 = 3.50</td>
<td>0.011</td>
</tr>
<tr>
<td>Autumnalis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,632; cases:949</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>Central Oklahoma</td>
<td>2,164</td>
<td>667.02</td>
<td>156 / 102.99 = 1.51</td>
<td>0.011</td>
</tr>
<tr>
<td>Grippotyphosa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,595; cases:795</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>Central Oklahoma</td>
<td>5,970</td>
<td>941.48</td>
<td>358 / 234.14 = 1.53</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pomona</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,646; cases:494</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>Central Kansas</td>
<td>2,358</td>
<td>675.28</td>
<td>115 / 59.58 = 1.93</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bratislava</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,678; cases:425</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>Central Indiana</td>
<td>919</td>
<td>249.73</td>
<td>44 / 18.69 = 2.35</td>
<td>0.006</td>
</tr>
<tr>
<td>Secondary cluster</td>
<td>North-central Colorado</td>
<td>65</td>
<td>4.86</td>
<td>10 / 1.32 = 7.56</td>
<td>0.020</td>
</tr>
<tr>
<td>Icterohaemorrhagiae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,699; cases:172</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>North-central Ohio</td>
<td>5,111</td>
<td>501.41</td>
<td>72 / 38.75 = 1.86</td>
<td>0.001</td>
</tr>
<tr>
<td>Canicola</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=18,698; cases:118</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary cluster</td>
<td>South-central Texas</td>
<td>116</td>
<td>22.42</td>
<td>12 / 0.57 = 21.21</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Example: FMD
Foot-and-mouth disease is a highly contagious, viral disease of domestic cloven-hoofed and many wild animals.
Worldwide occurrence of FMD
FMD in Korea

• Data (2010-2015): South Korea
  [link]

• Data (2010-2015): North Korea
  [link]
  http://www.oie.int/wahis_2/public/wahid.php/Wahid home/Home
  - Number of infected/susceptible animal at farm levels (with geographical information)
  - Outbreak date
### Discrete Poisson Model

- **Case file**
- **Population file**
- **Coordinate file**

<table>
<thead>
<tr>
<th>Location</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.40154</td>
<td>128.4522166</td>
</tr>
<tr>
<td>2</td>
<td>36.401681</td>
<td>128.453014</td>
</tr>
<tr>
<td>3</td>
<td>36.365585</td>
<td>128.41599</td>
</tr>
<tr>
<td>4</td>
<td>36.394074</td>
<td>128.471777</td>
</tr>
<tr>
<td>5</td>
<td>36.392993</td>
<td>128.474954</td>
</tr>
<tr>
<td>6</td>
<td>36.354015</td>
<td>128.423341</td>
</tr>
<tr>
<td>7</td>
<td>36.381978</td>
<td>128.434371</td>
</tr>
<tr>
<td>8</td>
<td>36.382545</td>
<td>128.436425</td>
</tr>
<tr>
<td>9</td>
<td>36.375359</td>
<td>128.422159</td>
</tr>
<tr>
<td>10</td>
<td>36.381929</td>
<td>128.430568</td>
</tr>
<tr>
<td>11</td>
<td>36.383693</td>
<td>128.421331</td>
</tr>
<tr>
<td>12</td>
<td>36.345966</td>
<td>128.412228</td>
</tr>
<tr>
<td>13</td>
<td>36.381998</td>
<td>128.435152</td>
</tr>
<tr>
<td>14</td>
<td>36.373215</td>
<td>128.424519</td>
</tr>
<tr>
<td>15</td>
<td>36.361887</td>
<td>128.412057</td>
</tr>
<tr>
<td>16</td>
<td>36.361214</td>
<td>128.433262</td>
</tr>
<tr>
<td>17</td>
<td>36.38209</td>
<td>128.435235</td>
</tr>
<tr>
<td>18</td>
<td>36.371644</td>
<td>128.441186</td>
</tr>
<tr>
<td>19</td>
<td>36.383089</td>
<td>128.423489</td>
</tr>
<tr>
<td>20</td>
<td>36.383074</td>
<td>128.431048</td>
</tr>
<tr>
<td>21</td>
<td>36.382128</td>
<td>128.472643</td>
</tr>
<tr>
<td>22</td>
<td>36.343786</td>
<td>128.391034</td>
</tr>
<tr>
<td>23</td>
<td>36.381494</td>
<td>128.454802</td>
</tr>
<tr>
<td>24</td>
<td>36.404506</td>
<td>128.423781</td>
</tr>
<tr>
<td>25</td>
<td>36.375689</td>
<td>128.453609</td>
</tr>
<tr>
<td>26</td>
<td>36.398099</td>
<td>128.415968</td>
</tr>
<tr>
<td>27</td>
<td>36.401094</td>
<td>128.421913</td>
</tr>
<tr>
<td>28</td>
<td>36.315029</td>
<td>128.474543</td>
</tr>
<tr>
<td>29</td>
<td>36.351712</td>
<td>128.289034</td>
</tr>
<tr>
<td>30</td>
<td>36.381163</td>
<td>128.555229</td>
</tr>
<tr>
<td>31</td>
<td>36.371847</td>
<td>128.394019</td>
</tr>
<tr>
<td>32</td>
<td>36.404778</td>
<td>129.032979</td>
</tr>
<tr>
<td>33</td>
<td>36.470858</td>
<td>128.374899</td>
</tr>
<tr>
<td>34</td>
<td>36.322111</td>
<td>128.230892</td>
</tr>
<tr>
<td>35</td>
<td>37.823</td>
<td>127.0082</td>
</tr>
<tr>
<td>36</td>
<td>38.0262</td>
<td>126.9561</td>
</tr>
<tr>
<td>37</td>
<td>37.8284</td>
<td>126.8316</td>
</tr>
<tr>
<td>38</td>
<td>39.8833</td>
<td>125.19</td>
</tr>
<tr>
<td>39</td>
<td>37.7133</td>
<td>126.6985</td>
</tr>
</tbody>
</table>
Results

1. Location IDs included:
   - Coordinates / radius: (37.224964 N, 128.081455 E) / 167.39 km
   - Time frame: 2014/12/1 to 2015/4/30
   - Population: 2172
   - Number of cases: 155767
   - Expected cases: 6523.33
   - Annual cases / 10000: 8923072.9
   - Observed / expected: 23.88
   - Relative risk: 254.51
   - Log likelihood ratio: 457679.09303
   - P-value: < 0.0000000000000001

2. Location IDs included:
   - Coordinates / radius: (38.733300 N, 126.128000 E) / 34.16 km
   - Time frame: 2010/12/1 to 2011/1/31
   - Population: 20
   - Number of cases: 6172
   - Expected cases: 73.77
   - Annual cases / 10000: 31265973.8
   - Observed / expected: 83.67
   - Relative risk: 86.76
   - Log likelihood ratio: 21334.34255
   - P-value: < 0.0000000000000001

3. Location IDs included:
   - Coordinates / radius: (37.000000 N, 128.500000 E) / 50% of the population size
   - Time frame: 2014/12/1 to 2015/4/30
   - Space and Time: starting from 50% to 10%

4. Location IDs included:
   - Coordinates / radius: (37.224964 N, 128.081455 E) / 167.39 km
   - Time frame: 2014/12/1 to 2015/4/30
   - Population: 2172
   - Number of cases: 155767
   - Expected cases: 6523.33
   - Annual cases / 10000: 8923072.9
   - Observed / expected: 23.88
   - Relative risk: 254.51
   - Log likelihood ratio: 457679.09303
   - P-value: < 0.0000000000000001

5. Location IDs included:
   - Coordinates / radius: (38.733300 N, 126.128000 E) / 34.16 km
   - Time frame: 2010/12/1 to 2011/1/31
   - Population: 20
   - Number of cases: 6172
   - Expected cases: 73.77
   - Annual cases / 10000: 31265973.8
   - Observed / expected: 83.67
   - Relative risk: 86.76
   - Log likelihood ratio: 21334.34255
   - P-value: < 0.0000000000000001

6. Location IDs included:
   - Coordinates / radius: (37.000000 N, 128.500000 E) / 50% of the population size
   - Time frame: 2014/12/1 to 2015/4/30
   - Space and Time: starting from 50% to 10%
ArcGIS – import (outbreak locations)
Outbreak locations (2010 Nov-2015 April)
ArcGIS – buffer (50%-50%)
Map with identified clusters
Time clusters (2010-2015)

**Number of cases**

- **Time frame:** 2014/12/1 to 2015/4/30
- **Coordinates / radii:** (27.224364 N, 120.881456 E) / 167.39 km
- **Number of cases:** 155767
- **Expected cases:** 6523.33
- **Annual cases / 100000:** 8923072.9
- **Observed / expected:** 23.88
- **Relative risk:** 254.51
- **Log likelihood ratio:** 457679.909303
- **P-value:** < 0.0000000000000001


- **Time frame:** 2010/12/1 to 2011/1/31
- **Coordinates / radii:** (27.224364 N, 120.881456 E) / 34.16 km
- **Number of cases:** 6172
- **Expected cases:** 73.77
- **Annual cases / 100000:** 31265473.8
- **Observed / expected:** 83.67
- **Relative risk:** 86.76
- **Log likelihood ratio:** 21334.34255
- **P-value:** < 0.0000000000000001
Exercises (50 mins)

- **Leptospirosis data**
  - Which model? (Bernoulli vs Poisson)
  - Create files (I already created for you)
  - Import data and Output
  - Determine: cluster size (from zero to 50%)
  - Comparison among different size (10%, 30% and 50%)

- **FMD data**
  - Which model? Space or time?
    (Poisson vs Space-time permutation)
  - Create files (You need to create files your own) - Notepad
  - Import data and Output
  - Determine: cluster size (from zero to 50%)
  - Comparison among different size (50%-50%, 50%-10%, 10%-50% and 10%-10%)
Poisson vs Permutation models (50%-50%)
Seasonality analysis
Introduction

• Procedures (seasonality using R)

Regional and Temporal Variations of Leptospira Seropositivity in Dogs in the United States, 2000–2010


Background: Previous studies have reported a seasonal increased risk for leptospirosis, but there is no consistent seasonality reported across regions in the United States.

Objectives: To evaluate and compare seasonal patterns in seropositivity for leptospirosis in dogs for 4 US regions (northeast [NE], midwest [MW], south-central [SC], and California-southern west coast [CS]).

Animals: Forty four thousand nine hundred and sixteen canine serum samples submitted to a commercial laboratory for microscopic agglutination tests (MAT) from 2000 through 2010.

Methods: In this retrospective study, positive cases were defined as MAT titers ≥1 : 3,200 for at least one of 7 tested serovars. Four geographic regions were defined, and MAT results were included in regional analyses based on hospital zip code.

A seasonal-trend decomposition method for times series was utilized for the analysis. Monthly variation in the seropositive rate was evaluated using a seasonal cycle subseries plot and logistic regression.

Results: Two thousand and twelve of 44,916 (4.48%) samples were seropositive. Compared to seropositive rates for February, significantly higher monthly rates occurred during the 2nd half of the year in the MW (OR 3.92–6.35) and NE (OR 2.03–4.80) regions, and only in January (OR 2.34) and December (OR 1.74) in the SC region. Monthly seropositive rates indicative of seasonality were observed earlier in the calendar year for both CS and SC regions.

Conclusions and Clinical Importance: Seasonal patterns for seropositivity to leptospires differed by geographic region. Although risk of infection in dogs can occur year round, knowledge of seasonal trends can assist veterinarians in formulating differential diagnoses and evaluation of exposure risk.

Key words: Dogs; Leptospira; Microscopic agglutination tests; Seasonal cycle subsseries plot; Seasonal-trend decomposition procedure based on loess; Seropositive.
Risk map

Rates of diagnoses of leptospirosis in the United States between 1970 and 2009 by university

Rate (per 100,000 dogs)

- 0.00 or unknown
- 0.01 - 27.46
- 27.47 - 50.54
- 50.55 - 78.91
- 78.92 - 115.59
- 115.60 - 140.01

0 250 500 1,000 Kilometers
What is STL and SCS?

• STL is a useful tool to visualize time series datasets that consists of:
  - Trend, seasonal and remainder components

• SCS helps to visualize patterns both between and with groups that consists of:
  - Horizontal lines: average for each month
  - Vertical lines: individual pattern for the same month in each year
Time series plot

Sero-positive rate (%)
STL plot

Trend

Seasonal

Remainder
Seasonal cycle subseries plot
<table>
<thead>
<tr>
<th>Month</th>
<th>Northeast</th>
<th>Mid-west</th>
<th>South central</th>
<th>California</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>1.30 (0.86-1.98)</td>
<td>1.18 (0.60-2.33)</td>
<td>1.32 (0.76-2.27)</td>
<td>Reference: 1</td>
</tr>
<tr>
<td>Feb</td>
<td>Reference: 1</td>
<td>Reference: 1</td>
<td>Reference: 1</td>
<td>1.93 (1.17-3.18)</td>
</tr>
<tr>
<td>Mar</td>
<td>0.71 (0.44-1.15)</td>
<td>1.26 (0.65-2.46)</td>
<td>1.08 (0.62-1.88)</td>
<td>1.98 (1.22-3.23)</td>
</tr>
<tr>
<td>Apr</td>
<td>0.91 (0.59-1.41)</td>
<td>0.87 (0.42-1.79)</td>
<td>1.28 (0.73-2.23)</td>
<td>1.60 (0.96-2.66)</td>
</tr>
<tr>
<td>May</td>
<td>1.06 (0.69-1.62)</td>
<td>1.70 (0.90-3.20)</td>
<td>1.45 (0.85-2.45)</td>
<td>1.49 (0.87-2.53)</td>
</tr>
<tr>
<td>Jun</td>
<td>0.90 (0.58-1.39)</td>
<td>1.16 (0.60-2.24)</td>
<td>1.08 (0.62-1.88)</td>
<td>1.22 (0.71-2.11)</td>
</tr>
<tr>
<td>Jul</td>
<td>1.23 (0.82-1.86)</td>
<td>1.42 (0.74-2.70)</td>
<td>1.35 (0.79-2.30)</td>
<td>0.95 (0.53-1.70)</td>
</tr>
<tr>
<td>Aug</td>
<td>1.61 (1.09-2.38)</td>
<td>1.59 (0.84-3.01)</td>
<td>1.07 (0.61-1.85)</td>
<td>1.02 (0.58-1.80)</td>
</tr>
<tr>
<td>Sep</td>
<td>2.04 (1.40-2.98)</td>
<td>1.59 (0.85-2.96)</td>
<td>1.30 (0.77-2.20)</td>
<td>0.76 (0.41-1.42)</td>
</tr>
<tr>
<td>Oct</td>
<td>2.29 (1.59-3.29)</td>
<td>1.72 (0.94-3.17)</td>
<td>1.30 (0.77-2.19)</td>
<td>0.62 (0.33-1.18)</td>
</tr>
<tr>
<td>Nov</td>
<td>3.40 (2.38-4.88)</td>
<td>2.31 (1.27-4.19)</td>
<td>1.47 (0.88-2.46)</td>
<td>1.60 (0.96-2.68)</td>
</tr>
<tr>
<td>Dec</td>
<td>2.39 (1.65-3.46)</td>
<td>2.00 (1.09-3.69)</td>
<td>1.74 (1.04-2.90)</td>
<td>1.31 (0.77-2.24)</td>
</tr>
</tbody>
</table>
Annual incidence rates for lepto and VE in humans

*Previous study showed that 17~71% of VE were caused by JE in Vietnam
Seasonality of VE in humans between 2004 and 2013 (Dien Bien, Hoa Binh, Lai Chau and Son La)

*Previous study showed that 17~71% of VE were caused by JE in Vietnam*
Exercises (15 mins)

• **VE monthly incidence rate from 2004 to 2013**
  - Open file: seasonality_VE and R_code_VE
  - Import into R:
  - Run the model
  - Interpretation
Time series analysis
Time series analysis of human and bovine brucellosis in South Korea from 2005 to 2010

Hu Suk Lee, Moon Her, Michael Levine, George E. Moore

Department of Comparative Pathobiology, College of Veterinary Medicine, Purdue University, 725 Harrison Street, West Lafayette, IN 47907, United States
OIE Reference Laboratory for Brucellosis, Bacterial Disease Division, Department of Animal and Plant Health Research, Animal, Plant and Fisheries Quarantine and Inspection Agency (QIA), 175 Anyang-ro, Manan-gu, Anyang-si, Gyeonggi 430-757, South Korea
Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 18 July 2012
Received in revised form 12 November 2012
Accepted 1 December 2012

Keywords:
Brucellosis
Time series
Lag
ARIMAX model
NBR model

ABSTRACT

Brucellosis is considered to be one of the most important zoonotic diseases in the world, affecting underdeveloped and developing countries. The primary purpose of brucellosis control is to prevent the spread of disease from animals (typically ruminants) to humans. The main objective of this study was to retrospectively develop an appropriate time series model for cattle-to-human transmission in South Korea using data from independent national surveillance systems. Monthly case counts for cattle and people as well as national population data were available for 2005–2010. The temporal relationship was evaluated using an autoregressive integrated moving average with exogenous input (ARIMAX) model [notated as ARIMA(p, d, q) – AR(p)] and a negative binomial regression (NBR) model.

Human incidence rate was highly correlated to cattle incidence rate in the same month and the previous month (both r = 0.82). In the final models, ARIMA(0, 1, 1) – AR(0, 1) was determined as the best fit with 191.5% error in the validation phase, whereas the best NBR model including lags (0, 1 months) for the cattle incidence rate yielded a 131.9% error in the...
Brucellosis

Transmission to Humans

Most transmission in South Korea
Introduction of project

• Rationale
  – Most human cases are related to not wearing protection, such as gloves and protective clothing when in contact with suspected cattle or materials

• Objective
  – To develop an appropriate time series model for cattle and humans in South Korea using data from independent national surveillance systems
Materials

• Study period: Jan 1, 2005 ~ Dec 31, 2010
• Data sources
  – Data collected on a yearly & quarterly basis:
    • KOSIS: national total population for human & cattle
      (Korean Statistical Information Service)

  – Data collected on a monthly basis:
    • KCDC: number of human cases
      (Korea Centers for Disease Control and Prevention)
    • AIMS: number of cattle cases
      (Animal Infectious Disease Data Management System)
Human & cattle incidence rates were calculated on a monthly basis – Cases / national total population

ARIMAX (autoregressive integrated moving average with exogenous input) model
- Cross correlation between human and cattle incidence rates
- Value of P < 0.05 was considered significant

Model validation
- Divided into model: construction (2005-2007) and validation (2008-2010) phases

The ARIMAX model is an extension of ARIMA (autoregressive integrated moving average) model:

\[ \hat{y}_t = \text{ARIMA} \left[ \text{Constant} + Y_{t-1} + \phi(Y_{t-1} - Y_{t-2}) - \theta e_{t-1} \right] + \text{AR} \left[ \beta_1 X_{t-1} + \beta X_{t-2} \right] \]
Methods

• Conducted simulation intervention scenarios (50% and 75% reductions in cattle cases)

• $\text{MAPE} = \left( \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{\text{Actual cases} - \text{predicted cases}}{\text{Actual cases}} \right| \right)$

  *Mean Absolute Percentage Error

• Software programs:
  – Microsoft Excel (Redmond, WA, USA)
  – STATA version 11.2 (Stata Corp., College Station, TX, USA)
Results
Monthly incidence rates of brucellosis between 2005 and 2010 in both human and cattle
## Descriptive statistics

<table>
<thead>
<tr>
<th>Month</th>
<th>Human cases</th>
<th>Human (% of total)</th>
<th>Cattle cases</th>
<th>Cattle (% of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>44</td>
<td>7.50</td>
<td>5,150</td>
<td>6.91</td>
</tr>
<tr>
<td>Feb</td>
<td>25</td>
<td>4.26</td>
<td>5,184</td>
<td>6.96</td>
</tr>
<tr>
<td>Mar</td>
<td>47</td>
<td>8.01</td>
<td>8,764</td>
<td>11.76</td>
</tr>
<tr>
<td>Apr</td>
<td>61</td>
<td>10.39</td>
<td>6,878</td>
<td>9.23</td>
</tr>
<tr>
<td>May</td>
<td>55</td>
<td>9.37</td>
<td>7,460</td>
<td>10.01</td>
</tr>
<tr>
<td>Jun</td>
<td>53</td>
<td>9.03</td>
<td>6,032</td>
<td>8.10</td>
</tr>
<tr>
<td>Jul</td>
<td>65</td>
<td>11.07</td>
<td>6,676</td>
<td>8.96</td>
</tr>
<tr>
<td>Aug</td>
<td>56</td>
<td>9.54</td>
<td>7,427</td>
<td>9.97</td>
</tr>
<tr>
<td>Sep</td>
<td>66</td>
<td>11.24</td>
<td>6,431</td>
<td>8.63</td>
</tr>
<tr>
<td>Oct</td>
<td>43</td>
<td>7.33</td>
<td>5,725</td>
<td>7.69</td>
</tr>
<tr>
<td>Nov</td>
<td>36</td>
<td>6.13</td>
<td>4,345</td>
<td>5.83</td>
</tr>
<tr>
<td>Dec</td>
<td>36</td>
<td>6.13</td>
<td>4,421</td>
<td>5.93</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>587</strong></td>
<td><strong>100</strong></td>
<td><strong>74,493</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

Cases of human and cattle brucellosis reported to the KCDC and AIMS between Jan 1 2005 and Dec 31 2010
Cross-correlation results

human and cattle incidence rates between 2005 and 2010 on a monthly basis with lags of 0-12 months

<table>
<thead>
<tr>
<th>Lags* (months)</th>
<th>Correlation (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.82</td>
</tr>
<tr>
<td>1</td>
<td>0.82</td>
</tr>
<tr>
<td>2</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>0.74</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>0.57</td>
</tr>
<tr>
<td>6</td>
<td>0.52</td>
</tr>
<tr>
<td>7</td>
<td>0.47</td>
</tr>
<tr>
<td>8</td>
<td>0.44</td>
</tr>
<tr>
<td>9</td>
<td>0.41</td>
</tr>
<tr>
<td>10</td>
<td>0.40</td>
</tr>
<tr>
<td>11</td>
<td>0.35</td>
</tr>
<tr>
<td>12</td>
<td>0.40</td>
</tr>
</tbody>
</table>

* The time periods between two observations. For example, lag 1 is between $Y_t$ and $Y_{t-1}$. Lag 2 is between $Y_t$ and $Y_{t-2}$. Time series can also be lagged forward, $Y_t$ and $Y_{t+1}$. 
### ARIMAX best model (First difference)\(^a\) : ARIMA (0, 1, 1) – AR (0, 1)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lags(months)</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>P-value</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unpredictable factors</td>
<td>1</td>
<td>-0.86</td>
<td>0.10</td>
<td>&lt;0.001</td>
<td>65.86%</td>
</tr>
<tr>
<td>Cattle incidence rate</td>
<td>0</td>
<td>1.84 x 10^{-4}</td>
<td>0.5 x 10^{-5}</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Cattle incidence rate</td>
<td>1</td>
<td>1.85 x 10^{-4}</td>
<td>0.5 x 10^{-5}</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>N/A</td>
<td>8.13 x 10^{-11}</td>
<td>1.66 x 10^{-9}</td>
<td>0.961</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)The first difference of \(Y_t = Y_{t-1} - Y_{t-2}\) (\(Y_t\) = human incidence rate at time \(t\))

\[
\hat{Y}_t = [ 8.13 \times 10^{-11} + Y_{t-1} + 1.84 \times 10^{-4} X_t + 1.85 \times 10^{-4} X_{t-1} + 0.86 e_{t-1} ] \times Y_{popt}
\]

\(\hat{Y}_t\) = the predicted number of human cases at time \(t\), \(X_t\) = the cattle incidence rate at time \(t\), 
\(e_{t-1}\) = unpredictable factors at time \(t-1\), \(Y_{popt}\) = the human population at time \(t\)
Examples

• Model Calculations
  – Actual human cases: 30
    \[26.12 = (8.13 \times 10^{-11} + 6.00 \times 10^{-7} + 1.84 \times 10^{-4} \times 1.28 \times 10^{-3} + 1.85 \times 10^{-4} \times 1.33 \times 10^{-3} + 0.86 \times e_{Aug2006}) \times 48,372,000\]

  – Actual human cases: 5
  – Predicted human cases in July 2010: 2.12
    \[3.95 = (8.13 \times 10^{-11} + 1.01 \times 10^{-7} + 1.84 \times 10^{-4} \times 1.30 \times 10^{-4} + 1.85 \times 10^{-4} \times 0.15 \times 10^{-4} + 0.86 \times e_{Jun2010}) \times 49,410,000\]
Predicted versus actual human cases

- Actual human cases
- ARIMAX model
- ARIMAX model - validation phase
Limitations

• A limitation of the study potentially lies in not being able to utilize data at the individual province level

• Exposure history of human brucellosis was not available
  – Restricting our ability to adjust for delayed recognition

• Other potential risk factors were not taken into account in the model
  – Surveillance of the wildlife has not been actively implemented to date
  – Farm levels and / or environmental factors
Thank you