Clustering US Tornadoes

Thomas H. Jagger, Florida State University
June 30, 2015, Aalborg, Denmark

Examining the Environmental Characteristics of Tornado Outbreaks in the United States with Spatial Clustering.

Integrating R-INLA with R Spatial Packages and the fpc, flexible procedures for clustering package.

What is a tornado outbreak?

- Tornadoes are common in the United States particularly in the Midwest and South.
  - Violent circulation attached to a parent cloud with rotational winds in excess 30 m/s.
  - Causes fatalities and complete destruction of buildings.
- Many tornadoes often occur in a single day known as an outbreak.
  - Define an outbreak as a single day with more than N tornadoes.
  - N=16
- Outbreak may be split geographically into separate regions.
  - Reflect local nature of outbreak.
  - Tornado characteristics and environment vary from region to region.
## Enhanced Fujita Scale

<table>
<thead>
<tr>
<th>EF Rating</th>
<th>Wind Speeds</th>
<th>Expected Damage</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-0</td>
<td>65-85 mph</td>
<td>‘Minor’ damage: shingles blown off or parts of a roof peeled off, damage to gutters/siding, branches broken off trees, shallow rooted trees toppled.</td>
</tr>
<tr>
<td>EF-1</td>
<td>86-110 mph</td>
<td>‘Moderate’ damage: more significant roof damage, windows broken, exterior doors damaged or lost, mobile homes overturned or badly damaged.</td>
</tr>
<tr>
<td>EF-2</td>
<td>111-135 mph</td>
<td>‘Considerable’ damage: roofs torn off well constructed homes, homes shifted off their foundation, mobile homes completely destroyed, large trees snapped or uprooted, cars can be tossed.</td>
</tr>
<tr>
<td>EF-3</td>
<td>136-165 mph</td>
<td>‘Severe’ damage: entire stories of well constructed homes destroyed, significant damage done to large buildings, homes with weak foundations can be blown away, trees begin to lose their bark.</td>
</tr>
<tr>
<td>EF-4</td>
<td>166-200 mph</td>
<td>‘Extreme’ damage: Well constructed homes are leveled, cars are thrown significant distances, top story exterior walls of masonry buildings would likely collapse.</td>
</tr>
<tr>
<td>EF-5</td>
<td>&gt; 200 mph</td>
<td>‘Massive/incredible’ damage: Well constructed homes are swept away, steel-reinforced concrete structures are critically damaged, high-rise buildings sustain severe structural damage, trees are usually completely debarked, stripped of branches and snapped.</td>
</tr>
</tbody>
</table>
Research Interests

- How does spatial clustering help us define the notion of a tornado outbreak?
- How do tornado outbreaks differ from each other?
- What mesoscale environmental conditions effect the frequency and energy of each tornado in an outbreak?
- What environment characteristics are common between tornado outbreaks?
- Could this method be used to identify conditions that lead to outbreaks?
Strategy for analysis

- Reduce to outbreak days and split the data set by days.
- For each day cluster tornadoes into groups.
- For each group find the convex hull of tornado start locations.
- Generate summary statistics of
  - tornadoes within each group.
  - environmental conditions within each convex hull.
- Model the relationship between the tornado statistics and environmental conditions.
Tornado Data Set

We use a modified tornado data set keeping tornado paths in the Midwest and South from 1979 to 2010 of at least EF0 (F0) strength. The data set is put into R SpatialLinesDataFrame objects or arrays. We use the R ggplot2 package for plotting data sets.

- Tornado data set is a spatial line data set with attributes from the SPC.
  - Use staring location and storm strength in our study.

- Reanalysis (Environmental) data from Climate Forecast System Reanalysis from NCAR
  - Initially we examine the CAPE and HLCY.
  - 1/2° by 1/2° resolution spatially
  - 4 times per day at 0000Z, 0600Z, 1200Z, 1800Z
Study Area
Clustering

- why? separate tornadoes into groups for analysis
  - Each group may have different characteristics

- what? The staring spatial locations for each tornado.
  - $x$ and $y$ values in the Lambert Conformal Conic projection
    - centered at $33^\circ N$
  - could use storm strength, EF magnitude, alone or as another clustering variable.

- Why use partitioning around medoid type of clustering?
  - Provide a sample observation representative of the whole.
  - Cluster around an actual tornado, not just an empty center.
  - The mediod tornado is not unique, just representative sample from cluster.
Clustering Example

- Subset tornado database to fit within our bounding box, and remove tornadoes without EF classification.
- Remove all days with less than 16 tornadoes.
- Split data by day and run medoid clustering algorithm.
- Create a convex hull around each cluster, enlarge it by 25km and convert to lon lat projection.

```r
xx = subset(TornC.spdf, Date == "2007-05-05")
cc = coordinates(xx)
best = pamk(cc, krange = 1:(N-1), alpha = .01)
cluster = best$pamobject$clustering
clustloc = split(1:length(xx),cluster)
Hulls = lapply(clustloc,function(i)
  spTransform(gBuffer(gConvexHull(xx[i,]),id=cluster[i[1]],width=25000),longlat))
Hulls.df = do.call("rbind",lapply(Hulls,fortify))
Map = get_map(location = c(lon=-99.5,lat=39.8), source = "google",
              matype = "roadmap", zoom = 6, color = "bw")
ggmap(Map, extent = "panel") + geom_point(aes(x = slon, y = slat),
                                           data = TornC.df[TornC.df$Date == dd, ],color = "black") +
gem_polygon(aes(x = long, y = lat, fill=id ,alpha=.5),
            data = Hulls.df,
            show_guide=FALSE)+scale_fill_manual(values=c("red","orange"))
```
Clusters on May 5, 2007
Summary Statistics and Analysis

- We calculate summary statistics within each cluster using the
  - tornado data set, for each group of tornadoes.
  - environmental data within each convex hull at 1200Z and 1800Z.
- We combine these into a single dataframe for analysis.
- We use R-INLA to analyze the relationship of
  - total kinetic energy and tornado counts to
    - Convective Available Potential Energy (CAPE),
    - Storm relative helicity (HLCY).
Summary Statistics for Tornadoes

- Total count of:
  - \( nT \) Tornadoes, at least EF0.
  - \( nST \) Strong Tornadoes, at least EF3.
- Total Kinetic Energy for all tornadoes.
  - \( \text{TKE} = \text{Height} \times \text{Area} \times \text{TKE per m}^3 \).
    - Height approximated as 1km
    - Area approximated by ellipsoid.
    - Fixed proportion of areas assigned to each EF strength.
    - Uses midpoint of EF scale \( E = \frac{1}{2} \rho V^2 \)
    - \( \rho \sim 10^3 \text{kg/m}^3 \)

Total kinetic energy in megajoules per m\(^3\) based on the tornado's strength is:

<table>
<thead>
<tr>
<th>EF</th>
<th>EF0</th>
<th>EF1</th>
<th>EF2</th>
<th>EF3</th>
<th>EF4</th>
<th>EF5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.570</td>
<td>0.661</td>
<td>0.786</td>
<td>0.919</td>
<td>0.974</td>
<td>1.054</td>
</tr>
</tbody>
</table>
Distribution of log10(TKE)
Environmental Conditions

- Thunderstorms may form if there is the potential for convection with lots of CAPE and little CIN.
  - Convective Available Potential Energy
  - Convective INhibition
  - CIN is required to get lots of CAPE.
- Thunderstorms may become super cell thunderstorms.
  - Updraft sustained by wind shear.
  - Storm rotates with directional wind shear.
- Super Cell thunderstorms may produce tornadoes.
  - Surface inflow wind contains rotation that upscale.
  - 0 - 3000m, total column storm relative helicity (HLCY)
- CAPE and HLCY measured in Joules/Kg or m² / s².
Sample Cape and Resulting Tornadoes
Summary Environmental Statistics

- Calculated the mean, maximum, median and standard deviation.
  - Within the convex hulls generated for each cluster
  - CAPE and HLCY at 1200Z and 1800Z
- Used only the 1800Z weighted mean values.
  - Values within each CAPE/HLCY grid box assumed to be the same value.
  - Grid intersection areas used as weights.
- Use rgeos functions.
  - glIntersects to find which grids are in each cluster.
  - glIntersection to find the spatial intersection of each grid to the cluster.
  - gArea to find the area of the intersection.
  - gConvexHull to create convex hull for each cluster.
  - gBuffer to expand each hull by 80km.
- Need to explore wind shear
  - Reanalysis data exists (u,v) for many levels from surface to stratosphere.
Analysis using R-INLA

- See http://www.r-inla.org, Bayesian modelling using integrated nested Laplace approximation.
- Previous work using INLA: Rpubs Tornado Climatology
- All covariates and response require scaling.
- Negative binomial distribution for counts.
- Gamma distribution for mean TKE per tornado in Terajoules, mTKET.
- Model covariates for mTKET are
  - LnT Logarithm for Number of Tornadoes,
  - CAPEK CAPE in K Joules,
  - HLCYH Storm relative helicity in H Joules.
- Model covariates for nT and nST are
  - CAPEK and HLCYH.

\[
\log(\mu) = \beta_0 + \beta_1 \text{LnT} + \beta_2 \text{CAPEK} + \beta_3 \text{HLCYH}
\]
Model Output for mTKET

Fixed effects:

<table>
<thead>
<tr>
<th>Term</th>
<th>mean</th>
<th>sd</th>
<th>0.025quant</th>
<th>0.5quant</th>
<th>0.975quant</th>
<th>mode</th>
<th>kld</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.690</td>
<td>0.234</td>
<td>-3.143</td>
<td>-2.693</td>
<td>-2.223</td>
<td>-2.697</td>
<td>0</td>
</tr>
<tr>
<td>lnT</td>
<td>0.782</td>
<td>0.069</td>
<td>0.643</td>
<td>0.783</td>
<td>0.916</td>
<td>0.784</td>
<td>0</td>
</tr>
<tr>
<td>CAPEK</td>
<td>-0.432</td>
<td>0.096</td>
<td>-0.618</td>
<td>-0.433</td>
<td>-0.242</td>
<td>-0.435</td>
<td>0</td>
</tr>
<tr>
<td>HLCYH</td>
<td>0.405</td>
<td>0.077</td>
<td>0.256</td>
<td>0.404</td>
<td>0.557</td>
<td>0.403</td>
<td>0</td>
</tr>
</tbody>
</table>

The model has no random effects

Model hyperparameters:

<table>
<thead>
<tr>
<th>Term</th>
<th>mean</th>
<th>sd</th>
<th>0.025quant</th>
<th>0.5quant</th>
<th>0.975quant</th>
<th>mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision parameter for the Gamma observations</td>
<td>0.563</td>
<td>0.026</td>
<td>0.513</td>
<td>0.563</td>
<td>0.975</td>
<td>0.563</td>
</tr>
<tr>
<td>Precision parameter for the Gamma observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expected number of effective parameters (std dev): 4.01(0.00)
Number of equivalent replicates: 158.27

Deviance Information Criterion: 825.34
Effective number of parameters: 4.65
## Model Outputs for nT

### Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>0.025quant</th>
<th>0.5quant</th>
<th>0.975quant</th>
<th>mode</th>
<th>kld</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.760</td>
<td>0.084</td>
<td>2.595</td>
<td>2.760</td>
<td>2.926</td>
<td>2.760</td>
<td>0</td>
</tr>
<tr>
<td>HLCYH</td>
<td>0.173</td>
<td>0.036</td>
<td>0.102</td>
<td>0.173</td>
<td>0.245</td>
<td>0.173</td>
<td>0</td>
</tr>
<tr>
<td>CAPEK</td>
<td>0.066</td>
<td>0.051</td>
<td>-0.034</td>
<td>0.066</td>
<td>0.167</td>
<td>0.066</td>
<td>0</td>
</tr>
</tbody>
</table>

The model has no random effects

### Model hyperparameters:

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>0.025quant</th>
<th>0.5quant</th>
<th>0.975quant</th>
<th>mode</th>
<th>kld</th>
</tr>
</thead>
<tbody>
<tr>
<td>size for the nbinomial observations (overdispersion)</td>
<td>2.63</td>
<td>0.166</td>
<td>2.32</td>
<td></td>
<td>0.5quant</td>
<td>0.975quant</td>
<td>mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expected number of effective parameters (std dev): 3.02(0.001)
Number of equivalent replicates: 209.61

Deviance Information Criterion: 4999.01
Effective number of parameters: 3.67
Discussion

- 1800Z mean cape and mean helicity significantly related to mean TKE, controlling for the number of tornadoes.
  - leads to -35% and 50% in posterior mean TKE for each 1000 J/Kg increase in CAPE and 100 J/Kg helicity.
  - Using the logarithm of nT, negates needing to model both mean TKE and total TKE.
- Helicity is strongly related to number of storms and strong storms per cluster.
  - Cape is marginally related.
  - leads to 7% and 19% in posterior mean nT and 16% and 83% in posterior mean nST respectively with each 1000 Joule increase in cape and 100 J helicity, controlling for the other covariate.
Summary

- Using cluster methods we can separate groups for further study.
  - The **fpc** package with the **pamk()** function was used for medoid clustering.
  - The clustering algorithm runs quickly, so is suitable for data sets in which you may want to perform many clusters.
    - We had over 500 cluster days with 634 clusters.
    - We had (400, 84, 7, 5, 3, 1, 1) days with (1, 2, 3, 4, 5, 7, 8) clusters respectively.
- Interesting findings within relationships inside clusters:
  - While CAPE is required for storms to form, the observed cape within clusters seems to be negatively related to TKE.
  - Increasing HLCY seems to increase both the number of tornadoes, and the mean TKE a measure of efficiency of tornado production.
Future Research

- Better identification of tornadoes and tornado clusters.
  - Outlier detection and removal.
- Better selection of geographical areas associated with each region of an outbreak.
  - Non convex regions possibly defined by level sets of tornado density estimates.
- Addition of other variables.
  - Storm shear in the environment.
  - Storm size in clustering algorithm.
Thank you for your time.

Analysis and Talk on http://rpubs.com/thjagger/

Thomas Jagger tjagger@fsu.edu