Toward Predictive Crime Analysis via Social Media, Big Data, and GIS

Anthony J. Corso
Agenda

• Introduction
  - Background
  - Objectives
  - Problem

• Data

• Artifact Outcomes

• Current and Future Research

• Conclusion
Introduction

- Objectives
  - Realize links between crime and domain specific data
  - Investigate Spatial and Temporal Analysis of Crime, Nearest Neighbor Clustering, Kernel Density Estimation, and Risk Terrain Modeling techniques
  - Investigate linguistic analysis of social media combined with crime and domain specific data for real-time predictive capabilities

- Problem
  - Identify predictive real-time ??? via social media and historical record
Theory Base for Research

- Abstract
- Concrete
- Existence
- Relational

Tweet Statements

Nominal
Ordinal
Interval
Ratio

Tweet Anatomy

"@MorganEastwood Evidently, your dad was going to be on the original Batman. What happened "I done been shot and had my head caved in so don’t expect me to break up no fights ...."",

"#PARANORMAL #FANTASY\n@FrostFyre\nWYCHFAE\nFun, Action-Packed Romance!!\nhttp://t.co/9c..."
Tweet Grammar
Natural Language Processing
Big Data

- What does it mean???
  - Millions, billions, trillions of records
  - A lot of data, A lot of data, A lot of data, A lot of data
  - Not, A lot of data is still small data

- Problem
  - Data pipe
    - JavaScript Object Notation (JSON)
    - General Architecture for Text Engineering (GATE)
    - Java
    - ArcGIS
Data, Maps, and GIS Overview

Small Scale

Large Scale
Data, Maps, and GIS Integration

Tweet Corpus  |  Crime Data  |  SNAP Locations
Hot Spot Maps Preliminary
Retrospective Hot Spot Maps

Hot Spot No Social Media

Hot Spot With Social Media
Regression Analysis

### Summary of OLS Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust SE</th>
<th>Robust t</th>
<th>Robust Pr [b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.703064</td>
<td>0.019981</td>
<td>235.374965</td>
<td>0.000000*</td>
<td>0.020054</td>
<td>234.515788</td>
<td>0.000000*</td>
</tr>
<tr>
<td>WORDNET</td>
<td>0.104344</td>
<td>0.054382</td>
<td>1.918726</td>
<td>0.055029</td>
<td>0.053274</td>
<td>1.958633</td>
<td>0.050163</td>
</tr>
</tbody>
</table>

### OLS Diagnostics

- **Input Features:** crimeTweetJoin
- **Number of Observations:** 20000
- **Multiple R-Squared [d]:** 0.000184
- **Adjusted R-Squared [d]:** 0.000134
- **Akaike's Information Criterion (AICc) [d]:** 95412.149378

### CRIMETYPE

- **Prob(>F), (1,19999) degrees of freedom:** 0.055033
- **Prob(>chi-squared), (1) degrees of freedom:** 0.050156
- **Prob(>chi-squared), (2) degrees of freedom:** 0.001995*
Contributions and Outcomes

- Predictive Crime Analysis via Social Media
  - Hypothesis: A GIS hot spot map or risk terrain model increases in precision and accuracy as a social media corpus is operationalization into GIS risk layers.
  - Independent Variable: Each different mapping treatment to produce a visualization

  - STAC: Spatial and Temporal Analysis of Crime: geographically locates the densest clusters of incidents on the map
  - Nnh: Nearest Neighbor Clustering: technique is based on a threshold distance to which the crime incidents are compared to identify clusters, e.g., city block
  - KDE: Kernel Density Estimation: continuous smoothed surface with variation of the density of crime over a study area
  - RTM: Two factors that have different operational spatial influences but can be spatially joined. The spatial overlap of risk factors creates a more risky environment where crime would be expected to occur in the future.
Current Research College Crime

Number of Colleges

- Seattle
- Salt Lake
- San Diego
- Dallas
- Chicago
- New Madrid
- New Orleans
- Boston

Legend:
- 7 - 9
- 10 - 12
- 13 - 16
Future Research Prediction

Tweet

Tweet Linguistic Decay
Conclusion and Questions

- Integrating domain specific and crime data
- A predictive social media artifact is possible
- GIS RTM artifact construction is encouraging