Combining GIS and Statistics: Data Visualization to Communicate Findings

Kristen Hocutt
Lu Zhang

March 7, 2019
Washington Statistical Society
WASHINGTON, DC
Everything happens somewhere

- Lakers Playoff Game @Staples Center
- Power Outage @Eugene area locations
- HR Incident @Scranton, PA
- Conference @Phoenix City Center
- Armed Protester @Seattle Distribution Center
- Machine down @San Francisco P&DC

- New Starbucks @Downtown Phoenix
- Amazon Prime Route Disruption @Reno & Irvine Locations
- Nearest Food truck Landing @ Salt Lake International Airport
- Winter Weather @Washington DC
- Cyber attack @Washington DC
- Route Disruption @San Francisco P&DC
- Winter Weather @Seattle Distribution Center
What are Spatial Statistics?
Spatial Statistics are a set of exploratory techniques for describing and modeling spatial distributions, patterns, processes, and relationships.
Spreadsheets

Data or Information?
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>4131 021 030916</td>
<td>4111</td>
<td>21</td>
<td>1567 F CAPS</td>
<td>3/9/2015</td>
<td>Coldstream</td>
<td>Homestead Mont Northease</td>
<td>14 (39.3222)</td>
</tr>
<tr>
<td>4139 026 030916</td>
<td>4119</td>
<td>20</td>
<td>2738 HARIH</td>
<td>3/9/2015</td>
<td>Coldstream</td>
<td>Homestead Mont Northease</td>
<td>14 (39.3212)</td>
</tr>
<tr>
<td>0244 034 031418</td>
<td>244</td>
<td>34</td>
<td>1708 W PF</td>
<td>3/14/2018</td>
<td>Union Square</td>
<td>Southern</td>
<td>9 (39.2854)</td>
</tr>
<tr>
<td>1125 027 031418</td>
<td>1125</td>
<td>27</td>
<td>1526 N CA</td>
<td>3/14/2018</td>
<td>Oliver</td>
<td>Eastern</td>
<td>12 (39.3080)</td>
</tr>
<tr>
<td>1127 051 030118</td>
<td>1127</td>
<td>51</td>
<td>1516 N BR</td>
<td>3/10/2018</td>
<td>Oliver</td>
<td>Eastern</td>
<td>12 (39.3078)</td>
</tr>
<tr>
<td>1138 114 031418</td>
<td>1138</td>
<td>154</td>
<td>1424 N BE</td>
<td>3/14/2018</td>
<td>Oliver</td>
<td>Eastern</td>
<td>12 (39.3069)</td>
</tr>
<tr>
<td>1145 032 031418</td>
<td>1145</td>
<td>32</td>
<td>1336 A45Q</td>
<td>3/14/2018</td>
<td>Oliver</td>
<td>Eastern</td>
<td>12 (39.3058)</td>
</tr>
<tr>
<td>1140B010 031418</td>
<td>1149B</td>
<td>10</td>
<td>1308 N CA</td>
<td>3/14/2018</td>
<td>Oliver</td>
<td>Eastern</td>
<td>12 (39.3054)</td>
</tr>
<tr>
<td>7056A018 030816</td>
<td>7056A</td>
<td>18</td>
<td>8418 6TH</td>
<td>3/8/2016</td>
<td>BROOKLYN</td>
<td>SOUther</td>
<td>10 (39.2395)</td>
</tr>
<tr>
<td>1488 010 030118</td>
<td>1488</td>
<td>10</td>
<td>2718 E OL</td>
<td>3/10/2018</td>
<td>Berea</td>
<td>Eastern</td>
<td>13 (39.3082)</td>
</tr>
<tr>
<td>1499 030 031418</td>
<td>1499</td>
<td>30</td>
<td>3008 E HAO</td>
<td>3/14/2018</td>
<td>Broadway East</td>
<td>Eastern</td>
<td>13 (39.3068)</td>
</tr>
<tr>
<td>1504 014 031518</td>
<td>1504</td>
<td>34</td>
<td>2407 E OL</td>
<td>3/15/2018</td>
<td>Broadway East</td>
<td>Eastern</td>
<td>13 (39.3077)</td>
</tr>
<tr>
<td>1505 055 031518</td>
<td>1505</td>
<td>55</td>
<td>1417 N MI</td>
<td>3/10/2018</td>
<td>Berea</td>
<td>Eastern</td>
<td>13 (39.3075)</td>
</tr>
<tr>
<td>1521 037 030118</td>
<td>1521</td>
<td>37</td>
<td>7225 E HO</td>
<td>3/10/2018</td>
<td>Berea</td>
<td>Eastern</td>
<td>13 (39.3067)</td>
</tr>
<tr>
<td>1542 076 032118</td>
<td>1542</td>
<td>76</td>
<td>1224 N PO</td>
<td>3/11/2018</td>
<td>Berea</td>
<td>Eastern</td>
<td>13 (39.3056)</td>
</tr>
<tr>
<td>1544 028 031818</td>
<td>1544</td>
<td>28</td>
<td>3113 E PRI</td>
<td>3/10/2018</td>
<td>Berea</td>
<td>Eastern</td>
<td>13 (39.3059)</td>
</tr>
<tr>
<td>1661 014 031318</td>
<td>1661</td>
<td>14</td>
<td>3027 MCE</td>
<td>3/13/2018</td>
<td>Elwood Park/Monument</td>
<td>Southeast</td>
<td>13 (39.2978)</td>
</tr>
<tr>
<td>0007 053 122315</td>
<td>7</td>
<td>53</td>
<td>1326 APRIL</td>
<td>12/29/2015</td>
<td>EASTERWOOD</td>
<td>WESTERN</td>
<td>7 (39.3084)</td>
</tr>
<tr>
<td>1667 014 031318</td>
<td>1667</td>
<td>14</td>
<td>429 N WAC</td>
<td>3/13/2018</td>
<td>CARE</td>
<td>Eastern</td>
<td>13 (39.2931)</td>
</tr>
<tr>
<td>1655 083 032118</td>
<td>1655</td>
<td>83</td>
<td>515 S POR</td>
<td>3/13/2018</td>
<td>McElroy Park</td>
<td>Southeast</td>
<td>12 (39.2972)</td>
</tr>
<tr>
<td>1656 063 031318</td>
<td>1656</td>
<td>63</td>
<td>512 N ROS</td>
<td>3/13/2018</td>
<td>McElroy Park</td>
<td>Southeast</td>
<td>12 (39.2972)</td>
</tr>
<tr>
<td>1657 062 031418</td>
<td>1657</td>
<td>62</td>
<td>514 N GLG</td>
<td>3/14/2018</td>
<td>McElroy Park</td>
<td>Southeast</td>
<td>12 (39.2973)</td>
</tr>
<tr>
<td>1658 074 031318</td>
<td>1658</td>
<td>74</td>
<td>53 N BEL</td>
<td>3/13/2018</td>
<td>McElroy Park</td>
<td>Southeast</td>
<td>12 (39.2977)</td>
</tr>
<tr>
<td>1659 002 031318</td>
<td>1659</td>
<td>52</td>
<td>503 N KEN</td>
<td>3/13/2018</td>
<td>McElroy Park</td>
<td>Southeast</td>
<td>12 (39.2970)</td>
</tr>
<tr>
<td>4581 010 030918</td>
<td>4581 010F</td>
<td>5312</td>
<td>2232 SPP</td>
<td>3/9/2018</td>
<td>Central Park Heights</td>
<td>Northwest</td>
<td>6 (39.3472)</td>
</tr>
<tr>
<td>4609 055 031418</td>
<td>4609</td>
<td>55</td>
<td>3701 MAN</td>
<td>3/14/2018</td>
<td>Central Park Heights</td>
<td>Northwest</td>
<td>6 (39.3430)</td>
</tr>
<tr>
<td>4616 005 031318</td>
<td>4616</td>
<td>5</td>
<td>3008 WOC</td>
<td>3/13/2018</td>
<td>Central Park Heights</td>
<td>Northwest</td>
<td>6 (39.3455)</td>
</tr>
<tr>
<td>5902A004 031518</td>
<td>5902A</td>
<td>4</td>
<td>3070 LAKE</td>
<td>3/15/2018</td>
<td>Belair-Edison</td>
<td>Northeast</td>
<td>13 (39.3208)</td>
</tr>
<tr>
<td>5907 001 032118</td>
<td>5907</td>
<td>31</td>
<td>3044 CHES</td>
<td>3/15/2018</td>
<td>Belair-Edison</td>
<td>Northeast</td>
<td>13 (39.3295)</td>
</tr>
<tr>
<td>5921 061 031518</td>
<td>5921</td>
<td>61</td>
<td>4340 SHAB</td>
<td>3/15/2018</td>
<td>Belair-Edison</td>
<td>Northeast</td>
<td>13 (39.3263)</td>
</tr>
<tr>
<td>0006 032 011509</td>
<td>6</td>
<td>32</td>
<td>2002 PRES</td>
<td>11/5/2009</td>
<td>EASTERWOOD</td>
<td>WESTERN</td>
<td>7 (39.3080)</td>
</tr>
<tr>
<td>0031 057 032316</td>
<td>31</td>
<td>57</td>
<td>1834 LAUF</td>
<td>3/23/2018</td>
<td>SANDTOWN-WINCHESTER</td>
<td>WESTERN</td>
<td>7 (39.3032)</td>
</tr>
<tr>
<td>0055C063 032316</td>
<td>0055C</td>
<td>63</td>
<td>1104 N CA</td>
<td>3/23/2018</td>
<td>SANDTOWN-WINCHESTER</td>
<td>WESTERN</td>
<td>7 (39.3031)</td>
</tr>
<tr>
<td>0073 024 031616</td>
<td>73</td>
<td>24</td>
<td>935 N MO</td>
<td>3/16/2018</td>
<td>SANDTOWN-WINCHESTER</td>
<td>WESTERN</td>
<td>7 (39.2988)</td>
</tr>
<tr>
<td>0075 016 032416</td>
<td>75</td>
<td>16</td>
<td>1515 MD</td>
<td>3/24/2016</td>
<td>SANDTOWN-WINCHESTER</td>
<td>WESTERN</td>
<td>7 (39.2977)</td>
</tr>
<tr>
<td>0082 058 032416</td>
<td>82</td>
<td>58</td>
<td>805 N BRH</td>
<td>3/24/2016</td>
<td>MIDTOWN-EDMONTSON</td>
<td>WESTERN</td>
<td>7 (39.2978)</td>
</tr>
<tr>
<td>0096 052 031516</td>
<td>96</td>
<td>52</td>
<td>707 N MO</td>
<td>3/15/2017</td>
<td>HARLEM PARK</td>
<td>WESTERN</td>
<td>9 (39.2968)</td>
</tr>
<tr>
<td>1109 940 030118</td>
<td>1109</td>
<td>140</td>
<td>1704 N BR</td>
<td>3/10/2018</td>
<td>Oliver</td>
<td>Eastern</td>
<td>12 (39.3097)</td>
</tr>
<tr>
<td>0149 944 021616</td>
<td>149</td>
<td>44</td>
<td>310 N BRL</td>
<td>3/16/2016</td>
<td>FRANKLIN SQUARE</td>
<td>WESTERN</td>
<td>9 (39.2922)</td>
</tr>
<tr>
<td>0244 074 031518</td>
<td>244</td>
<td>76</td>
<td>1708 LEMFR</td>
<td>3/15/2016</td>
<td>SOUTHERN PARK</td>
<td>SOUTHERN</td>
<td>9 (39.2858)</td>
</tr>
</tbody>
</table>
When you look at a spreadsheet...

<table>
<thead>
<tr>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
You ask for more

- Mean
- Standard Deviations
- Min and Max
- ...

1
2
3
4
5
Maps

Data or Information?
Same goes for maps!
We can do more
Reconnecting Milwaukee

A BikeAble™ Study of Opportunity, Equity and Connectivity
ArcGIS Connection with Open Source Libraries
Navigating the Python Ecosystem

Leveraging the Python Ecosystem

Enterprise & Online

Jupyter Notebook

ArcGIS API for Python

ArcPy

Desktop

PyTorch

CONDA

SciPy

Requests

pandas

TensorFlow

Keras

matplotlib

SymPy

OpenCV

NumPy

dmlc

XGBoost
Sample Use Case:

Developing a Weather Repository
• R: A widely used statistical programming language
• More than 10,000 Packages
• Expand Workflows
Sample Use Cases:

- Predicting Seagrass Locations

- Ingesting / Analyzing Multidimensional Scientific Data

- Modeling Vacancy Rates in DC -- from Baltimore Vacancy Data
Clustering, Prediction & Classification

Kristen Hocutt
Subjectivity of Maps
The map as data
High Priority 911 Calls in Baltimore
The map as data
High Priority 911 Calls in Baltimore

Where are the hot spots? Where is the variation greater?
The map as data
High Priority 911 Calls in Baltimore

Where are the hot spots? Where is the variation greater?
The subjectivity of visual pattern analysis

Where are the hot spots? Where is the variation greater?
Minimizing the subjectivity
Turning the map into information
Complete Spatial RANDOMNESS
Is there a PATTERN?
What are the chances this happened RANDOMLY?22
z-scores
p-values
Hot Spot Analysis

given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic
...how do we know if it’s SIGNIFICANTLY different???
MATH!
Density-based Clustering

finds clusters based on feature locations
DBSCAN – defined distance

HDBSCAN – self adjusting

OPTICS – multi-scale
Spatial Analysis of Traffic Accidents

A Density Based Clustering Approach
Models

Representative generalizations used for prediction
Why model

Use information we have to predict information we don’t have

- Which areas are most contaminated?
- What drives sales?
- Which buildings will fail inspection?
- What will the weather be like tomorrow?
When we can’t trust a model

Mimics training dataset and models noise instead of generalizing a trend
Divorce Rate in Maine vs Per Capita Consumption of Margarine

Correlation: 0.992558
Number People Who Drowned by Falling into a Swimming-Pool vs Number of Nicolas Cage Films

Correlation: 0.666004
Many ways to model

- Generalized Linear Regression
- Geographically Weighted Regression
- Forest-based Classification and Regression
Generalized Linear Regression
Modeling linear relationships
Three model types

- Gaussian – continuous
- Logistic – binary
- Poisson – count
Gaussian

Ordinary Least Squares
Dependent Variable

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

What are you trying to predict or understand?
Exploratory Regression

Variables:
- Cultural
- Environmental
- Socio Economic
- Lifestyle
- Spatial

Tests all variable combinations for:
1. Redundancy
2. Completeness
3. Significance
4. Bias
5. Performance

Creates Output Diagnostic Report
Geographically Weighted Regression
Exploring spatial variation
each feature gets a separate equation
Forest-based Classification & Regression

Predicting using machine learning
Random subset of data and variables used in each tree
Majority vote wins
Potential Applications/Where it’s used Today

Crop Yield

Housing Values

Lead Poisoning

JIAXUAN YOU, XIAOCHENG LI, MELVIN LOW, DAVID B. LOBELL, STEFANO ERMON-STANFORD UNIVERSITY
Data Visualization

Kristen Hocutt
Why visualize data?

Convert slow reasoning tasks into fast perception tasks
<table>
<thead>
<tr>
<th>1399683</th>
<th>635200</th>
<th>2908933</th>
<th>2106392</th>
</tr>
</thead>
<tbody>
<tr>
<td>6962578</td>
<td>5401748</td>
<td>8275961</td>
<td>4885854</td>
</tr>
<tr>
<td>1019462</td>
<td>3922722</td>
<td>6074504</td>
<td>3018484</td>
</tr>
<tr>
<td>1345609</td>
<td>8875318</td>
<td>6636256</td>
<td>9978939</td>
</tr>
<tr>
<td>724027</td>
<td>5401748</td>
<td>3880520</td>
<td>4796559</td>
</tr>
<tr>
<td>844322</td>
<td>6581982</td>
<td>9913774</td>
<td>2995330</td>
</tr>
<tr>
<td>587106</td>
<td>2815039</td>
<td>6539407</td>
<td>4681639</td>
</tr>
<tr>
<td>5731663</td>
<td>2917750</td>
<td>26538203</td>
<td>19383475</td>
</tr>
<tr>
<td>4796559</td>
<td>38120066</td>
<td>6636256</td>
<td>9853722</td>
</tr>
<tr>
<td>1628760</td>
<td>1885932</td>
<td>5235100</td>
<td>742404</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>1399683</td>
<td>635200</td>
<td>2908933</td>
<td>2106392</td>
</tr>
<tr>
<td>6962578</td>
<td>5401748</td>
<td>8875318</td>
<td>8275961</td>
</tr>
<tr>
<td>1019462</td>
<td>3922722</td>
<td>6581982</td>
<td>6074504</td>
</tr>
<tr>
<td>1345609</td>
<td>1336265</td>
<td>2815039</td>
<td>6636256</td>
</tr>
<tr>
<td>3111844</td>
<td>2917750</td>
<td></td>
<td>3880520</td>
</tr>
<tr>
<td>6656872</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>724027</td>
<td>1877879</td>
<td>11575704</td>
<td>9913774</td>
</tr>
<tr>
<td>844322</td>
<td>19631599</td>
<td>12914651</td>
<td>6539407</td>
</tr>
<tr>
<td>587106</td>
<td>12817894</td>
<td>629120</td>
<td>26538203</td>
</tr>
<tr>
<td>5731663</td>
<td>3601157</td>
<td>927030</td>
<td>5903286</td>
</tr>
<tr>
<td>1628760</td>
<td>1050228</td>
<td>1885932</td>
<td>5235100</td>
</tr>
</tbody>
</table>

**Total:** 38120066
Good Viz vs Bad Viz
U.S. Smartphone Marketshare

- Apple: 39.0%
- Nokia: 19.5%
- Other: 19.5%
- Motorola: 7.4%
- Palm: 3.1%
- RIM: 21.2%
Visualizations to support spatial analysis

Distributions and frequency

Category comparisons

Relationships and correlations

Change over time or distance
When a map (alone) isn’t the best option...
When a map (alone) isn’t the best option...
When a map (alone) isn’t the best option...
When a map (alone) isn’t the best option...
Key Takeaways

Something happens Everywhere

Open Source paired with COTS can enhance workflows

Data Visualization & Messaging drive quick decisions
Resources

- Learn ArcGIS
- R Bridge
- ArcGIS Pro
- Spatial Statistics Page
Questions?

Kristen Hocutt- khocutt@esri.com
Lu Zhang- l.zhang@esri.com