What’s Special About GeoAI and Spatial Data Science?

Shashi Shekhar
McKnight Distinguished University Professor, Univ. of Minnesota
www.cs.umn.edu/~shekhar
Happy President’s Day!

We are grateful to many great presidents for transformative contributions such as Spatial Data and Geo-Intelligence.

… Eisenhower … Clinton …

1960: Eisenhower reviews photo from Satellite Tiros I.
Acknowledgements

• P.I., Connecting the Smart-City Paradigm with a Sustainable Urban Infrastructure Systems Framework to Advance Equity in Communities, NSF (1737633), $2.5 M, 9/1/2017 - 8/31/2021.

• P.I., Spatio-temporal Informatics for Transportation Science, NSF (1901099), $1.2M, 8/1/19-7/31/23.

• P.I., EAGER: Spatiotemporal Big Data Analysis to Understand COVID-19 Effects, $100K, NSF (2040459), 9/1/20-8/31/22.

• P.I., Identifying Aberration Patterns in Multi-attribute Trajectory Data with Gaps, $600K, USDOD-NGA (HM0476-20-1-0009), 6/15/20- 6/14/23.


A Geo-Intelligence and Spatial Data Science Story

1854: What causes Cholera?

Collect & Curate Data

Discover Patterns, Generate Hypothesis

Test Hypothesis (Experiments)

Develop Theory

- Miasma theory
- Water pump
- Impacts: Hygiene, Separate drinking water and sewage systems, ...

Q? What are Choleras of today?
Q? How may Geo-Intelligence and Spatial Data Sc. Help?
### What has changed?

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spatial Data Revolution</strong></td>
<td>Smaller data from surveys, few satellites and sensors</td>
<td><strong>Spatial Big Data</strong> from Nano-satellites, Billions of GPS enabled devices, …</td>
</tr>
<tr>
<td>Better and Platform</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial Processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial Data Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial Data Visualization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Large Constellations of Small Satellites

- Hi-frequency (e.g., daily or hourly) time-series of imagery of entire earth
- Large Constellations
  - 2021: Planet Labs: 200+ satellites: daily Earth scan (1m resolution, visible+NIR bands)

GPS is Ubiquitous Today!

- 2 billion GPS receivers in use, will hit 7 billion by 2022.
- Besides location, it reference time for critical infrastructure
  - Telecommunications industry
  - Banks
  - Airlines...

- GPS is the single point of failure for the entire modern economy.

- 50,000 incidents of deliberate (GPS) jamming last two years
  - Against Ubers, Waymo’s self-driving cars, delivery drones from Amazon

Spatial Data Revolution

- Remotely sensed Imagery
  - Thousands of (Nano-)satellites
  - UAVs, Aerial imagery, …

- (GPS-) Location traces
  - Billions of phones, vehicles, …
  - Spatio-temporal patterns of life

- Others
  - Vehicle On-board diagnostics
  - Geo-social media, …

- Why is it interesting?
  - See previously inconspicuous
  - Monitor hard to monitor areas
  - Solve previously unsolvable problems

Monitor Global Crops for Early Warning

- Last century: US Wetland inventory took 4 decades and $400M
- Now: Global crop-health maps produced monthly for early warning and action
Better Visibility of Activities at Oceans and Seas

• “For years it’s been impossible to see illegal acts happening at sea, from overfishing to human rights abuses. Now that’s changing” (Source (b))

• **Automatic Identification System (AIS):**
  – Ships (> 300 tons) report location
  – Collision Avoidance (augment marine radar)
  – Monitor fishing and cargo fleet
  – Search and Rescue, Statistics and Economics

• **Example:**
  – A fishing vessel switched off AIS for 15 days.
  – near the Galapagos Marine Reserve

Sources:  
(b) How to spot the secretive activities of rogue fishing boats, bbc.com, 7 June 2018.  
Monitor energy use and emissions => Eco-Routing


Oct. 2021: Google Maps supports Eco-routes
### What has changed?

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spatial Data Revolution</strong></td>
<td>Smaller data</td>
<td>Spatial Big Data</td>
</tr>
<tr>
<td><strong>Better Access and Platforms</strong></td>
<td>Smaller platforms, e.g., ESRI Arc/Info, Postgis</td>
<td>Big Compute, e.g., AWS Earth, ESRI GIS Tools for Hadoop,</td>
</tr>
</tbody>
</table>

**Source:** Original PC ARC/INFO Brochure, OCT. 22, 2010 / M. ARTZ, HTTPS://GISANDSCIENCE.COM/2010/10/22/ORIGINAL-PC-ARCINFO-BROCHURE/
Geo-AI Platforms

Global GEO-AI, e.g., Compare geo-policy alternatives

Future Platforms
- I/O Intensive
- Data Analytics Platforms
- Supercomputers
- Large Clusters
- Workstations w/ accelerators
- Embedded w/ accelerators
- Embedded/ Edge
- Future Platforms

Tiny GEO-AI
- 10ms – seconds
- Hours
- Days
- Weeks
- Months

Move closer to interactive?

Note: Accelerators include GPU, TPU, FPGA
Cloud Repositories expand Access to Spatial Big Data

- 2008: USGS gave away 35-year LandSat satellite imagery archive
  - Analog of public availability of GPS signal in late 1980s
- 2017: Cloud-based repositories of geospatial data
  - Explosion in machine learning on satellite imagery to map crops, water, buildings, roads, …

<table>
<thead>
<tr>
<th>Data Sources and Projects</th>
<th>Google Earth Engines</th>
<th>NEX</th>
<th>AWS Earth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation, Landsat, LOCA, MODIS, NAIP</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>NOAA</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>CHIRPS, GeoScience Australia, GMap, NASS, Oxford Map, PSDI, WHRC, WorldClim, WorldPop, WWF, BCCA, FLUXNET</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Earth on AWS
Build planetary-scale applications in the cloud with open geospatial data.
## What has changed?

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Data Revolution</td>
<td>Smaller Data</td>
<td>Spatial Big Data</td>
</tr>
<tr>
<td>Spatial Data Access and Platforms</td>
<td>Smaller platforms</td>
<td>Big Compute, e.g., Cloud (e.g., AWS Earth, ESRI GIS Tools for Hadoop)</td>
</tr>
<tr>
<td><strong>Spatial Data Processing</strong></td>
<td>Fairly manual, labor-intensive (Geo-Intelligence)</td>
<td>More automation (Geo-Augmented-Intelligence (Geo-AI))</td>
</tr>
<tr>
<td>Spatial Data Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial Data Visualization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Excerpt:
Geospatial imagery “software” “specially designed” for training a Deep Convolutional Neural Network to automate the analysis of geospatial imagery and point clouds, and having all of the following:

1. Provides a graphical user interface that enables the user to identify objects (e.g., vehicles, houses, etc.) from within geospatial imagery and point clouds in order to extract positive and negative samples of an object of interest;
2. Reduces pixel variation by performing scale, color, and rotational normalization on the positive samples;
3. Trains a Deep Convolutional Neural Network to detect the object of interest from the positive and negative samples; and
4. Identifies objects in geospatial imagery using the trained Deep Convolutional Neural Network by matching the rotational pattern from the positive samples with the rotational pattern of objects in the geospatial imagery.
Traditional AI vs. Deep Learning

• From Satellite Imagery: Classify Land-cover, Map buildings
• Ex. 2009 Haiti Earthquake: Map building damage [1]

---

Deep Learning for Geo-Object Detection

- Q: How many vehicles in a parking lot? City?
- Ex.: Estimate truck supply in a city (CH Robinson).
- **Old Computer Vision workflow**
  - Many steps, each adds error
- **New Deep Learning Workflow – fewer steps**
  - Aerial imagery (3 inch pixels, Twincites, MN, USA)
  - NAIP Imagery (1 meter pixels, 2017)
  - MA Buildings Data (https://www.cs.toronto.edu/~vmnih/data/)

- **Detected Geo-objects**
  - Cars, trucks, Houses, …
  - Method: Convolutional Neural Networks (YOLO)

**Note:** NAIP = National Agriculture Imaging Program (USDA)

Mapping Trees from Remote Sensing Imagery

• Why?: Protect Powerlines, Manage Emerald Ash Borer, Green infrastructure Equity
• Input: LiDAR + Remotely Sensed Imagery + (NAIP Ground Truth)
• Approach: Tree Inference by Minimizing Bound-and-band Errors (TIMBER)
  – Optimization to find tree locations and sizes
  – Deep learning constructs features separating trees & non-trees (e.g., light pole)

## What has changed?

<table>
<thead>
<tr>
<th></th>
<th>Last Century</th>
<th>This Century</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Data Revolution</td>
<td>Smaller Data</td>
<td>Spatial Big Data</td>
</tr>
<tr>
<td>Spatial Data Access and Platforms</td>
<td>Smaller platforms</td>
<td>Big Compute, e.g., Cloud (e.g., AWS Earth, ESRI GIS Tools for Hadoop)</td>
</tr>
<tr>
<td>Spatial Data Processing</td>
<td>Fairly manual, labor-intensive (Geo-Intelligence)</td>
<td>More automation Geo-Augmented-Intelligence (Geo-AI)</td>
</tr>
<tr>
<td><strong>Spatial Data Science</strong></td>
<td>“One-size fit all” AI applied to spatial data</td>
<td>Virtuous cycle between Geo and AI</td>
</tr>
</tbody>
</table>

![A Day in the Life of an Intelligence Analyst](image)

**AI:**
- One size does NOT fit all.

**Geo:**
- Foundational Breakthroughs
- Use-inspired Innovations
Why go beyond Object Detection?

Source: A. Karpathy (2012), The state of Computer Vision and AI: we are really, really far away
http://karpathy.github.io/2012/10/22/state-of-computer-vision/
Geo-AI beyond Computer Vision

• Research Initiatives
  – USDOE ORNL Trillion Pixel Challenge; Map all buildings in US and beyond
  – Self-driving: Detect road objects in adverse weather (e.g., rain, snow, dust)
  – 2017-20: DARPA Geospatial Cloud Analytics: Crop yield, fracking, illegal fishing
  – 2020-onwards: IARPA Space-based Machine Automated Recognition Technique (SMART)
    • Change (Construction Stage); Underground: (Subterranean Challenge)

• Text, multimedia, knowledge graphs
  – Question Answering Systems: Answer geographic questions,
  – Geo-locate a picture/video, Fact-check maps, …

• Points, Polygons, Trajectories
  – Characterize spatio-temporal patterns of life
Need for AI (Vision) to break out of “RGB+Lidar” box

A downpour in Las Vegas at the CES technology show a year and a half ago may prove to have been a watershed moment in the race to develop autonomous cars. While other companies promoting their experimental self-driving vehicles had to keep them parked in the rain, one company, AdaSky, demonstrated how its sensors could see people hundreds of feet ahead even in a downpour, and even when they were wearing white and standing against a white background.

“Thermal imaging is the best sensor at detecting people, day or night,” Chris Posch of FLIR Systems said.

These High-Tech Sensors May Be the Key to Autonomous Cars

The New York Times
Sept. 26, 2019
Human vision → Superhuman Vision

- Remote sensing use richer Sensors than AI (Computer Vision):
  - Ex. Electromagnetic radiation emitted by objects above absolute zero (0° Kelvin(K), −273°C)

Source: directthermography.co.uk

Source: imagine.gsfc.nasa.gov
Should AI learn Richer Spatial Concepts?

Q? What is distance between Washington D.C. and U.S.A.?
• Zero (Washington D.C. is inside U.S.A.)
• NSF Open Knowledge Networks initiative grants on geo-knowledge networks!

Google search results:
- Distance from Washington D.C. to U.S.A.
- Map showing travel time and distance:
  - 18 h 23 min (1,175.1 mi) via I-70 W
Spatial Data Types: OGC Simple Features Standard

- Spatial Concepts: Point, LineString, Polygon, Collections
- Relationships: Topological, Metric, …
- Helps feature selection for machine learning
  - Ex. Distance to key geo-features, Neighbor relationship

From Objects Detection to Pattern Mining

After object detection we look for Spatial patterns
- Hotspots, Spatial clusters
- Spatial outlier, discontinuities
- Co-locations, segregations
- Spatiotemporal predictions

Spatial Data Mining is
The process of discovering
- interesting, useful, non-trivial patterns
- from large spatial datasets

Spatial Pattern Mining Challenges

• Traditional pattern mining methods not robust in face of
  – Challenge 1: Noise
  – Challenge 2: Spatial continuity
  – Challenge 3: Auto-correlation, Heterogeneity, Edge-effect, …

Dealing with Noise & Chance Patterns

• **Statistics: Deal with Noise**
  – Quantify uncertainty, confidence, …
  – Is it (statistically) significant?
  – Is it different from a chance event or rest of dataset?
    • e.g., SaTScan finds circular hot-spots

• **Spatial Statistics, Spatial Data Mining**
  – Auto-correlation, Heterogeneity, Edge-effect, …
Robust Clustering (Hotspot Detection)

- **Problem definition**
  - **Inputs**: Collection of event locations, Test statistic; Significance level
  - **Output**: Significant clusters (hotspots)
  - **Constraints**: Avoid chance patterns despite non-trivial noise in data

Challenge 2: Continuous Space

- Traditional relationship mining methods not robust
  - Result changes if spatial partitioning changes
  - Similar to Gerrymandering risk, Formally, Modifiable Areal Unit Problem (MAUP)
  - Neighbor Graph Based Measures are more robust

### Details:

A Metric of Spatial Cross-Correlation

• Ripley’s Cross K-Function Definition

\[ K_{ij}(h) = \lambda_j^{-1} E \text{[number of type } j \text{ event within distance } h \text{ of a randomly chosen type } i \text{ event]} \]

• Interpretation
  • Which pairs are frequently co-located
  • Statistical significance
Co-locations

• **Given:** A collection of different types of spatial events
• **Find:** Co-located subsets of event types

Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).
Illustration of Cross-Correlation

- Illustration of Cross K-function for Example Data
Spatial Colocation

Feature set: (𝐸, 𝐿, 𝐻)

Feature Subsets:  

Participation ratio (pr):

\[ \text{pr}(\bullet, \begin{array}{c} \bullet \\bullet \end{array}) = \text{fraction of } \bullet \text{ instances neighboring feature } \{\bullet\} = \frac{2}{3} \]
\[ \text{pr}(\bullet, \begin{array}{c} \bullet \\bullet\end{array}) = \frac{1}{2} \]

Participation index (pi):

\[ \text{pi}(\begin{array}{c} \bullet \\bullet \end{array}) = \min\{ \text{pr}(\bullet, \begin{array}{c} \bullet \\bullet \end{array}), \text{pr}(\bullet, \begin{array}{c} \bullet \\bullet \end{array}) \} \]
\[ = \min(\frac{2}{3}, \frac{1}{2}) = \frac{1}{2} \]

Participation Index Properties:

1. Computational: Non-monotonically decreasing like support measure
2. Statistical: Upper bound on Ripley’s Cross-K function

Details: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H. Yan, H. Xiong).
Challenge 3: Spatial Auto-correlation

- Traditional Statistics, ML, Data Mining
- Ubiquitous i. i. d. assumption
  - Data samples independent of each other
  - From identical distribution
- Problem
  - Ignores auto-correlation, heterogeneity
  - Salt n Pepper noise
### Ex. Salt n Pepper Noise

**Input:**
- (a) aerial photo
- (b) aerial photo
- (c) true classes

**Output:**
- (d) DT prediction (Salt n Pepper Noise)
- (e) SDT prediction

**Training samples:** upper half  
**Test samples:** lower half  
**Spatial neighborhood:** maximum 11 pixels by 11 pixels

Traditional decision tree

Inputs: table of records

<table>
<thead>
<tr>
<th>ID</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$\Gamma$</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>green</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>3</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>green</td>
</tr>
<tr>
<td>K</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>red</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>green</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
<td>3</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>red</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>1</td>
<td>-1</td>
<td>green</td>
</tr>
<tr>
<td>J</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>red</td>
</tr>
<tr>
<td>L</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>P</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>Q</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>R</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>red</td>
</tr>
</tbody>
</table>

Spatial decision tree

Inputs:
- feature maps, class map
- Rook neighborhood

<table>
<thead>
<tr>
<th>ID</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$\Gamma$</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>green</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>3</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>green</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>green</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
<td>3</td>
<td>0.3</td>
<td>green</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>red</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>1</td>
<td>-1</td>
<td>green</td>
</tr>
<tr>
<td>J</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>red</td>
</tr>
<tr>
<td>L</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>P</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>Q</td>
<td>3</td>
<td>2</td>
<td>0.3</td>
<td>red</td>
</tr>
<tr>
<td>R</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>red</td>
</tr>
</tbody>
</table>

Feature $f_1$

$$I(f_1 \leq 1)$$

green red

Feature $f_2$

$$I(f_1 \leq 1) \times \Gamma$$

green red

Predicted map

Class map

Predicted map

<table>
<thead>
<tr>
<th>feature test</th>
<th>information gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1 \leq 1$</td>
<td>0.50</td>
</tr>
<tr>
<td>$f_2 \leq 1$</td>
<td>0.46</td>
</tr>
<tr>
<td>$f_2 \leq 2$</td>
<td>0.19</td>
</tr>
</tbody>
</table>
## Modeling Spatial Auto-correlation

- **Traditional Models**, e.g., Regression (with Logit or Probit),
  - Linear Regression (LR), Bayes Classifier, ...
- **Semi-Spatial**: LR with auto-corr. regularizer
  \[ \varepsilon = \| y - \beta X \|^2 + \| \beta X - \beta X_{\text{neighbor}} \|^2 \]
- **Spatial**
  - Spatial autoregressive model (SAR)
  - Markov random field (MRF) based Bayesian Classifier

<table>
<thead>
<tr>
<th>Traditional</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ y = X\beta + \varepsilon ]</td>
<td>[ y = \rho W y + X\beta + \varepsilon ]</td>
</tr>
<tr>
<td>[ \Pr(C_i \mid X) = \frac{\Pr(X \mid C_i) \Pr(C_i)}{\Pr(X)} ]</td>
<td>[ \Pr(c_i \mid X, C_N) = \frac{\Pr(C_i) \Pr(X, C_N \mid c_i)}{\Pr(X, C_N)} ]</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>Convolutional Neural Networks</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>Spatial Decision Trees</td>
</tr>
</tbody>
</table>
Spatial Auto-Regression & Parameter Estimation

<table>
<thead>
<tr>
<th>Name</th>
<th>Model</th>
<th>ρ: the spatial auto-regression (auto-correlation) parameter</th>
<th>W: n-by-n neighborhood matrix over spatial framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Linear Regression</td>
<td>[ y = \mathbf{x}\beta + \varepsilon ]</td>
<td>[ y = \rho \mathbf{W}y + \mathbf{x}\beta + \varepsilon ]</td>
<td></td>
</tr>
</tbody>
</table>

- **Maximum Likelihood Estimation**

\[
\ln(L) = \ln|\mathbf{I} - \rho \mathbf{W}| - \frac{n \ln(2\pi)}{2} - \frac{n \ln(\sigma^2)}{2} - SSE
\]

- Computing determinant of large matrix is a hard (open) problem!
  - size(W) is quadratic in number of locations/pixels.
  - Typical raster image has Millions of pixels
  - W is sparse but not banded.

**Details:** A parallel formulation of the spatial auto-regression model for mining large geo-spatial datasets, SIAM Intl. Workshop on High Perf. and Distr. Data Mining, 2004. (with B. Kazar)
Spatial Heterogeneity

- Knowledge of location can improve land-cover and object recognition
  - Q? Which pictures show snow?
    - (a) Salt Marsh (Runn of Kutch, Gujarat, India)
    - (b) Snow
    - (c) Snow

- Coarse Satellite Imagery (e.g., 30m pixels)
  - More effective for large mono-crop farms the small mixed-crop plots

- However, Convolutional Neural Networks does not model geographic heterogeneity.
Modeling Spatial Heterogeneity: GWR

- Geographically Weighted Regression (GWR)
  - Goal: Model spatially varying relationships
  - Example: $y = X\beta' + \varepsilon$
    Where $\beta'$ and $\varepsilon'$ are location dependent

Source: resources.arcgis.com
Spatial Variability Aware Neural Networks (SVANN)

- Each NN parameter is a map i.e., a function of location
  - Similar to Geographically Weighted Regression

A Neural Network (NN)

SVANN

- Evaluation:
  - Urban Garden Detection across Hennepin County, MN and Fulton County, GA.
  - SVANN outperformed OSFA by 14.34% on F1-scores.

## What has changed?

<table>
<thead>
<tr>
<th></th>
<th>Last Century</th>
<th>This Century</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spatial Data Revolution</strong></td>
<td>Smaller Data</td>
<td>Spatial Big Data</td>
</tr>
<tr>
<td><strong>Spatial Data Access and Platforms</strong></td>
<td>Smaller platforms</td>
<td>Big Compute, e.g., Cloud (e.g., AWS Earth, ESRI GIS Tools for Hadoop)</td>
</tr>
<tr>
<td><strong>Spatial Data Processing</strong></td>
<td>Fairly manual, labor-intensive (Geo-Intelligence)</td>
<td>More automation Geo-Augmented-Intelligence (Geo-AI)</td>
</tr>
<tr>
<td><strong>Spatial Data Science</strong></td>
<td>“One-size fit all” AI applied to spatial data</td>
<td>Virtuous cycle between Geo and AI</td>
</tr>
<tr>
<td><strong>Spatial Data Visualization</strong></td>
<td>Maps, albums</td>
<td>Spatio-temporal, 3D</td>
</tr>
</tbody>
</table>
Towards Time-Travel and Depth in Virtual Globes

- Virtual globes are snapshots
- How to add time? depth?
  - Ex. Google Timelapse: 260,000 CPU core-hours for global 30+frame video
  - https://earthengine.google.com/timelapse/
  - Dubai coastal expansion
  - Chicago O’Hare airport
  - Doha, Qatar
  - Marina Center, Singapore (Wikipedia entry)
  - Salt Lake, Bidhannagar, Kolkata, WB, India
  - UMN, Minneapolis; airport, MN, USA

http://googleblog.blogspot.com/2013/05/a-picture-of-earth-through-time.html
Data that are geographically referenced or contain some type of location markers are both common and of high value (e.g., data subject to state-specific policies, laws and regulations; demographic data from the census; location traces of smartphones and vehicles; remotely sensed imagery from satellites, aircraft and small unmanned aerial vehicles; volunteered geographic information; geographically referenced social media postings). A 2011 McKinsey Global Institute report estimates a value of “about $600 billion annually by 2020” from leveraging personal location data to reduce fuel waste, improve health outcomes, and better match products to consumer needs. Spatial data are critical for societal priorities such as national security, public health & safety, food, energy, water, smart cities, transportation, climate, weather, and the environment. For example, remotely-sensed satellite imagery is used to monitor not only weather and climate but also global crops for early warnings and planning to avoid food shortages.
However, spatial data presents unique data science challenges. Recent court cases that address gerrymandering, the manipulation of geographic boundaries to favor a political party, offer a high-profile example. Instances of such exploitation of the modifiable areal unit problem (or dilemma) is not limited to elections since the MAUP affects almost all traditional data science methods in which results (e.g., correlations) change dramatically by varying geographic boundaries of spatial partitions. The fundamental geographic qualities of spatial autocorrelation, which assumes properties of geographically proximate places to be similar, and geographic heterogeneity, where no two places on Earth are exactly alike, violate assumptions of sample independence and randomness that underlie many conventional statistical methods. Other spatial challenges include how to choose between a plurality of projections and coordinate systems and how to deal with the imprecision, inaccuracy, and uncertainty of location.

A UCGIS Call to Action: Bringing the Geospatial Perspective to Data Science Degrees and Curricula

University Consortium for Geographic Information Science

Summer 2018
Spatial Data Science Tools

measurements. To deal with such challenges, practitioners in many fields including agriculture, weather forecast, mining, and environmental science incorporate geospatial data science methods such as spatially-explicit models, spatial statistics, geo-statistics, geographic data mining, spatial databases, etc.

6 H. Miller and J. Han, Geographic Data Mining and Knowledge Discovery, CRC Press, 2009 (2nd Ed.).

A UCGIS Call to Action:
Bringing the Geospatial Perspective to Data Science Degrees and Curricula

University Consortium for Geographic Information Science

Summer 2018
Summary

• Spatial Data has already transformed our society
  • It is only a beginning!
  • It promises astonishing opportunities in coming decade

• AI has promise but faces major challenges
  – Rich Data Types, e.g., lineStrings, polygons, …
  – High cost of errors, Spatial Heterogeneity, …

• Ask
  – Sponsors: Nurture approaches to overcome challenges (Geo-AI)
  – Academics: Include Spatial topics in courses and curricula

UCGIS Call to Action:
Bringing the Geospatial Perspective to Data Science Degrees and Curricula

Summer 2018
References : Surveys, Overviews

- Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE Transactions on Knowledge and Data Mining, 29(10):2318-2331, June 2017.
- Parallel Processing over Spatial-Temporal Datasets from Geo, Bio, Climate and Social Science Communities: A Research Roadmap. IEEE BigData Congress 2017: 232-250.