Spatiotemporal Methodologies and Analytics for Extreme Weather Study – using Dust Storm Event as an Example

Manzhu Yu
NSF Spatiotemporal Innovation Center
Department of Geography and Geoinformation Science
George Mason University
Outline

- Research going on in the NSF Spatiotemporal Innovation Center (GMU site)
  - Spatiotemporal Computing Infrastructure
  - Climate Spark
  - Planetary Defense
  - Big Data & Deep Learning

- Extreme weather identification and tracking (Dust Storm)
NSF I/UCRC Spatiotemporal Innovation Center
www.stcenter.net

- NSF University, Industry and Government collaborative research center for spatiotemporal thinking, computing, and applications
  - Computing: GMU center for intelligent spatial computing (CISC)
  - Thinking: UCSB Center for Spatial Studies (CSS)
  - Applications: Harvard Center for Geographic Analysis (CGA)
- Industry advisory board (IAB)
Research Topics
Spatiotemporal Computing Infrastructure

- Website portal: [http://cloud.gmu.edu/](http://cloud.gmu.edu/)
- Cloud platform: [https://stc.dc2.gmu.edu/dc2us2/login.jsp](https://stc.dc2.gmu.edu/dc2us2/login.jsp)

Two servers. Each server contains 24 CPU cores, 32G Memory and 1TB disk.

Private Cloud:
- Eucalyptus: 4800 CPU Cores, 4800 GB RAM, 400TB Storage
- Openstack: 4272 CPU Cores, 4272 GB RAM, 200TB Storage
Spatiotemporal Computing Infrastructure

OS:
Ubuntu 14.04 x 252
CentOS 6.6 x 252

IBM dx360
Intel Xeon CPU X5660 @2.8Ghz x 24
24GB RAM
Total: 504 servers

GMU Network & Internet

Gateway Server

1Gbps

InfiniBand (20Gbps) x 2

InfiniBand Switches

InfiniBand (2x12x20Gbps)

InfiniBand Switches

Ethernet (1Gbps)

Gateway Server

1Gbps

10Gbps
ClimateSpark: Distributed Computing Framework for Big Climate Data Analytics

https://github.com/feihugis/ClimateSpark

- Ease the operation and visualize the result
- Data analytics in a high performance fashion
- Fast spatiotemporal query
- Big data storage

Spatiotemporal Index

Bridge the gap between the logical and the physical data model

Each leaf node:
- Logical data info
  - variable name
  - geospatial range
  - temporal range
  - chunk corner
  - chunk shape

- Physical data pointer
  - node list
  - fileId(byte offset, byte length, file name)

Geospatial join query

Soil type, Washington DC

+

Impervious type, Washington DC

Soil type under the landscape features
Planetary Defense

Knowledge Base
- Domain-specific knowledge base
- Ranking
- Recommendation
- Semantic reasoning

Analytics
- Name entity recognition (NER)
- Relation extraction (RE)
- Summarization
- Topic modeling
- Profile mining
- Link analysis

Data System
- Content Index
- Special Index
- Repositories
- Data management
- Data access
- Security
- Operation

Sources
- Web pages
- Documents
- Access logs

Gateway
Query

Internet
Crawling
Upload
Generate
Planetary Defense System architecture

Query-independent features
- Publish date
- Webpage PageRank score
- Domain PageRank score

Query-dependent features
- Title, URL, content, anchor text TF-IDF
- Topic modelling results
Automatically learn and detect disaster events from big data

- LANCE Rapid Response MODIS images
- Images of extreme weather events are downloaded
- Each class contains about 200 images

Hyper parameter optimization

- Classification accuracy
- Training time
- Learning rates
- Batch size
- Training epochs
- Image processing parameters
- Number of layers
- Convolutional filters
- Convolutional kernel size
- Dropout fractions

Dust    Fire    Hurricane    Plume

https://lance-modis.eosdis.nasa.gov/cgi-bin/imagery/gallery.cgi
Classifying extreme weather events using MODIS true color images

- **Objective:** Using deep learning techniques to detect extreme weather events
- **Tool:** TensorFlow
- We use **transfer learning**, starting with a model that has been already trained on another problem to solve a similar problem
- **Model:** Inception v3 network
  - Trained for the ImageNet Large Visual Recognition Challenge using the data from 2012
  - It can differentiate between 1,000 different classes, like Dalmatian or dishwasher
- We use the same network, but retrain it to classify a small number of classes: **dust, fire, hurricane, and plume**
Integrating the use case into the big data deep learning platform: Progress

1) Data preprocessing module
   • Data filter
     • randomize function (complete), normalize, resample functions

2) Providing a common workflow of building TensorFlow model, using CNN as an example

3) Explainable classification
   • Visual explanation (using Local Interpretable Model-agnostic Explanations, LIME). Explain why this data is classified into a certain class
   • Future: semantic explanation, e.g. “This is dust event, because it has mesoscale coverage of yellowish airborne dust”

https://github.com/marcotcr/lime
Classifying tweets into disaster response themes

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caution and Advice (CA)</td>
<td>Warning or a piece of advice given about a related incident</td>
<td>Flooded neighborhoods in Norfolk and its approaching low tide.</td>
</tr>
<tr>
<td>Casualties and Damage (CD)</td>
<td>Information about casualties or infrastructure damage</td>
<td>This tree and power lines are down at the corner of Station Road and Bethlehem Pike in Quakertown.</td>
</tr>
<tr>
<td>Information Sources (IS)</td>
<td>A message from an official news source, media or government</td>
<td>@NYCMayorsOffice: Mayor: All @NYCSchools are closed tomorrow.</td>
</tr>
<tr>
<td>People</td>
<td>People found or missing</td>
<td>PLEASE RT: IF ANYONE HAS ANY INFO ON THE WHEREABOUTS OF AMANDA LANZONE OF FAR ROCKAWAY, PLEASE PASS IT ON TO @Vicosuave89</td>
</tr>
<tr>
<td>Donation and Aid (DA)</td>
<td>Goods or services offered or needed by victims</td>
<td>I don't have any money to donate but I have lots of time, where can I help/volunteer in #Hoboken? Who do I call?</td>
</tr>
</tbody>
</table>

- **Objective:**
- Deep learning based methods to classify tweets into different disaster relief themes
- Facilitates rapid tweet identification for disaster response purposes

A simple CNN model for tweet topic classification

- The first step is preprocessing, where each word of the tweet is represented by an integer.
- The preprocessed tweet passes through the first layer, word embedding, which expands the word integers to a larger matrix and represents them in a more meaningful way.
- The convolution layer then extracts features from the word embedding and transforms them through global max pooling.
- Then two fully connected layers predict the themes of each tweet.
- Dropout layers are utilized before the convolution layer and the last fully connected layer.
- Activation functions are used after the convolution layer and the fully connected layers.
• A significant increase of tweet number for “Caution and Advice” can be observed on October 29, since the wind, rain, and flooding occurred in the city during that night.
• An increase for the class “People” on October 30, and a continuous increase of “Casualties and Damage” during the two days of October 30 and 31.
• Moving forward, we observe a gradual increase for “Donation and Aid” throughout the study time period until it reaches its peak on Nov 3, and decreases gradually for the rest of the time.
Twitter dataset

- Most tweets for “Caution and Advice” and “People” are from the communities of lower Manhattan, since news reports broadcasted that this area would be impacted and drew people’s attention.

- Tweets about “Casualties and Damage” are more distributed in the area indicating damages of storm surge and high winds occurred throughout the area.

- Similar patterns can be observed for the class “Donation and Aid” mentioning about “red cross”, “FEMA”, and “volunteering”
Results

- The classification accuracy on the training data and test data changes over time.

- The accuracy rises gradually towards 1.0, whereas the test accuracy reaches ~0.81.

- This indicates that our network is overfitting:
  - the network is memorizing the training set, without understanding texts well enough to generalize to the test set.

- As a major problem in neural networks, overfitting is difficult to address especially when deep learning networks often have very large numbers of weights and biases.

- In this case, the network has 2,138,155 parameters with 289,255 trainable parameters.
Results

• Although techniques like dropout and regularization have been utilized in our network, the sign of overfitting is still not improving.

• The reason is that our training dataset is relatively small with 1151 samples, comparing to other benchmarking large scale datasets, e.g. AG’s news: 120,000 train samples and Amazon Review Full: 3,600,000 train samples.

• The size of our train and test data is limited by the nature of twitter data, which was harvested real time through Twitter Streaming API.

• We are extending the dataset, integrating from multiple hurricane disasters to increase the dataset will produce better performance with this CNN model.
Comparative studies among the CNN classifier, a SVM classifier, and a Logistic Regression classifier

- True positive rate (Recall)
- Positive predictive value (Precision)
- F1-score: harmonic average of the precision and recall

- Precision: the CNN model had values over 0.81 while the SVM model had 0.72 and LR had 0.56.
- Almost similar behavior is observed in the Recall and in F1-score.
- These findings state clearly that CNN outperforms traditional text mining approaches for tweet classification presenting potential for further development on tweet theme identification.
Web log mining

• The application of data mining techniques to discover interesting usage patterns from Web log data in order to better serve the needs of Web-based applications.

• 201402 PODAAC logs: 3.52GB, 67,27,710 (~7 million) lines in total, it took more than one hours to finish the whole process one virtual machine (8 CPU cores, 16G memory).

<table>
<thead>
<tr>
<th>Podaac.log.201401</th>
<th>Import log</th>
<th>Crawler detection</th>
<th>Session identification</th>
<th>Total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>3140</td>
<td>130</td>
<td>603</td>
<td>3873</td>
</tr>
</tbody>
</table>

• The amount of logs continues to grow up with time. It is hard to store them in one physical machine.

• Leveraging cloud computing and a Hadoop-Elasticsearch based framework to speed-up the log mining process.
Framework

Log files into HDFS from various sources

Index logs into Elasticsearch using Spark

Analyze logs using Elasticsearch & Spark

Master Node

Worker Node

Worker Node

Virtual Machines

Computing platform

Hybrid Cloud Computing
Comparison of “ocean OR wind” search results

Not Relevant !! (SST or SSH altimeter datasets)
- MUDROD results:
  - Recall similar
  - Precision improved!


https://mudrod.jpl.nasa.gov/#/
Spatiotemporal Methodologies and Analytics for Extreme Weather Study – using Dust Storm Event as an Example

Manzhu Yu
NSF Spatiotemporal Innovation Center
Department of Geography and Geoinformation Science
George Mason University
Overview

• Dust research and relationship with GIScience
• Detecting dust events from 4D simulation data
  • 3D dust identification
  • Movement tracking
  • Spatiotemporal data framework
• Conclusions – Future directions
A view of dust storm events

Dust storm is a common phenomenon in arid and semi-arid regions, often arising when strong surface wind uplifts fine-grained dust particles into the air.

Phoenix Dust Storm a "100-Year Event“, 2011, July 5th
Source: Youtube
A view of dust process

- Dust emission
- Turbulent diffusion and vertical advection
- Horizontal advection
- Sedimentation, dry and wet deposition

Atmospheric dust process
Source: WMO
A view of dust research

In order to mitigate the hazardous impact of dust storms, it is crucial to detect an upcoming dust storm and predict its impact and uncertainty level.

Benefits:
- **Scientists:**
  - Observe and better understand the evolution and transportation of dust storm over space and time
- **Policy-makers:**
  - Obtain early information and design mitigation plans
- **General public:**
  - Obtain warning information and take relevant responses
Scientific Questions

Observations

Understanding of dust processes

Dust prediction

Modeling

Mitigation

Where and when exactly does a dust event happen?

How do dust events transport in a regional and global scale?
Feature identification

Where and when exactly does a dust event happen?
Background

- Currently, manual interpretation and simple visualization of 2D/3D maps are generated to analyze 4D dust model results

- **Limitations** of current methods:
  - Only experts and scientists can interpret
  - 2D: cannot represent real dust storm objects in the 3D real world
  - 3D: cannot capture the movement patterns of dust events
  - Current operational models produce a data volume of TB daily, manual interpretation is no longer adequate.

- Automated dust storm feature discovery should be conducted through more sophisticated analytical and data mining methods
Dust Identification

• Why identify?
  • detect the presence, origin, direction, and speed of dust storms

Challenges:

A. How to define a storm cell?
  • Single threshold VS Multiple thresholds
  • Heuristics

B. How to identify individual storm cells?
  • False merger problem
  • Cluster of storms
Specification of dust features

• A contiguous 3D volume
  • dust concentration value exceeds a certain threshold
  • affected surface area exceeds a certain size

• Affected surface area: $10^3$ km$^2$
  • statistics of peak dust storm process derived from observational data (Lei and Wang, 2008)

• Dust concentration threshold:
  • Combine the set of automatic generated multi-thresholds with a standard set of multi-thresholds
    • Standard set: Barcelona Supercomputing Center (i.e., 20, 40, 80, 160, 320, 640, 1280, 2650 µg/m$^3$)
    • Automatic technique: Otsu’s multi-threshold approach (Liao et al. 2001)
    • Ostu $\rightarrow$ Standard set
Example with multi-threshold (20, 40, 80 μg/m³)

(a) 2D identified dust storm cores on surface level, threshold 20 μg/m³
(b) 2D identified dust storm cores on surface level, threshold 40 μg/m³
(c) Cluster (Result a) and core (Result b) connection
(d) Splitting result of threshold pair (20, 40 μg/m³)
(e) 2D identified dust storm cores on surface level, threshold 80 μg/m³
(f) Cluster (Result d) and core (Result e) connection
(g) Final identification result on surface level

Cluster 1 has multiple cores [1, 2]
Cluster 4 has a single core [3]
Cluster 5 has a single core [4]
Cluster 6 has multiple cores [5, 6, 7]
Cluster 7 has a single core [8]
Cluster 8 has multiple cores [5, 6, 7]
Cluster 9 has a single core [7]
Cluster 10 has a single core [8]

Identifying dust storm objects

- Grouping contiguous sequences of grid cells for which the dust concentration value exceeds a selected threshold
- Based on Region Growing (Zucker 1976)
- Extend to 3D process
- Generate different regions representing different core or cluster

Connecting cores with clusters

- Three possibilities:
  - A cluster contains no cores
  - A cluster contains only one core
  - A cluster contains two or more cores
- Needs further splitting to solve false merger problem and cluster of storms
Why further splitting?

• Geospatial objects tend to interact with other objects while largely keeping their own properties

• If considering interacting objects as one large system, the significance of feature tracking and the study of a feature’s life cycle is decreased

Detecting false merger

• Found in experiment:
  • Weak connection does not exist through all vertical levels of the 3D dust storm feature
  • Once a breach at a particular vertical level is detected, dust storm feature is likely to be a false merger
Splitting clusters with multiple cores

- Grow multiple cores at the same time in 6 directions
- Check grid cells that belong to multiple cores
- Determine which core this grid cell belongs to by applying a spatial-intensity constraint, inspired by Bankman et al. (1997)
  - Calculate slopes between the local maxima of different cores and the grid cell
  - Assign the grid cell to the core with the largest slope value (the fastest route from the edge of the core to the local maxima)

(a) Splitting with spatial constraint
(b) Splitting without spatial constraint
(c) Original dust concentration
Indications of results

• 3D dust feature identification
  • Address the false merger problem and isolate substorms within a cluster
  • Benefit the further efforts of dust storm feature tracking
  • Facilitate the auto-processing of simulation datasets, further feature mining

• Appropriate for other geospatial feature identification from 3D simulations:
  • thunderstorm, jet streams, and ocean objects

• Spatial identification → Spatiotemporal detection
  • Track evolutionary stages of dust events, movement patterns, transport paths, etc.
Event tracking:

How do dust events transport in a regional and global scale?
Movement tracking

• Geospatial phenomena, such as those studied in meteorology, oceanography, and geosciences:
  • intrinsically spatiotemporal (3D: latitude, longitude, and time, or 4D: latitude, longitude, vertical level, and time) in nature
  • and highly dynamic

• Simulations become too complex for researchers to analyze manually:
  • when and where events happen
  • how long an event lasts
  • how the event evolves

• Tracking benefits:
  • Automatic testing hypothesis and refine assumptions
  • Discovering and understanding the complex pattern over long time-period and over large dataset
Challenges of tracking dust events

• Highly involve *vertical* dimension
  • Dust up-lifting from arid and semi-arid regions
  • Transporting in the air
  • Depositing back to the ground

• Features tend to evolve and interact, while 2D objects in computer vision interact less frequently.

• As pointed out by Wilson et al. (1998), the key reason for poor extrapolation forecasts is not errors in forecast displacement, but the *growth and decay of storms* in the forecast period.

• An important aspect of storm-tracking algorithms is how they handle the *splitting* or *merging* of storms. (Lakshmanan and Smith, 2010)
Tracking procedure

• Using the identification results:
  • Unique ID
  • Attributes: volume, intensity-weighted centroid, and the centroid’s corresponding attributes, such as weed speed, pressure, temperature, as so on
• Boundaries

• Assumption: dust storm objects from a later time step have partial overlap with those from an earlier time step

• Possibilities of overlap:
  • Continuation, Merge, Split, Appear, Disappear

• After tracking applied:
  • full lifecycles, trajectories of dust storms to identify their movements
  • Moving direction, speed, growth/decay rate, will be calculated based on the centroid of each dust storm object

Generating long-term transport pattern of dust events?
A spatiotemporal data framework for dust event tracking

**ST_Object**: object that continues to exist through its lifecycle

**ST_Relation**: record the split and merge relations

**ST_Event**: An event may consist of several ST_Objects, which interact with each other

**Trajectory**, **CoverageSeries**, **VolumeSeries**: record the movement of each event in different dimensionalities

Advantages of this data framework

• Establish the links between different times of the same spatiotemporal object
  • Easy to maintain consistent representations after updates and precludes topological queries along this dimension
  • Finding a moving object involves a brute force search, and checking if two objects at different scales are equivalent can now be inferred directly

• The evolving geometry of an event can be efficiently retrieved from the framework in multiple dimensions, i.e. trajectory (1D), coverage (2D), and volume (3D) changing in time.
Reconstructing the four types of entities from dust feature identification and tracking

Tracking the movement of a single event
Seasonal analysis of reconstructed dust events

Querying dust events originating from Libya desert in the four seasons of 2014
Evaluation of identified dust storm based on visibility observation

- Identification threshold: mean value of the multi-thresholds
- Visibility data:
  - station-based weather observation, in unit of meters
  - dusty conditions are defined as the present visibility observation of less than 10 km
- Evaluation result: mean POD: 85%, mean FAR: 12%
Evaluation of tracked dust events with NASA Earth Observatory

(a) West Africa Event Feb 28, 2014

(b) Oman Event Jun 15, 2014
Indications of result

• A scientific framework
  • conceptually model the natural phenomenon as an event
  • introduce identification and movement tracking approach to reconstruct events by searching through 4D simulation datasets
  • analyze the evolutions and dynamic movements of the events

• Advancements of the data framework:
  • direct handling of the evolution of natural phenomena
    • trajectory, and coverage and volume dynamics
    • enhancing the query and analysis of different dimensionalities for various purposes
  • introducing a workflow of extracting dust events from 4D simulation datasets through a feature identification and tracking approach
Conclusions

1. Identification of dust storm features to detect from simulations (IJGIS)

2. Movement tracking (Computer and Geosciences) and spatiotemporal data framework
   - Complex transport patterns
Future directions

1. Spatiotemporal statistic analysis
   • Discover the relationship between detected movement (seasonal/annual/inter-annual transport pattern) with possible impacting factors
   • Rainfall, land use, atmospheric cycle, soil moisture, ENSO...

2. Big data + Deep Learning:
   • Utilize deep learning techniques to extract geospatial events, including dust events, hurricane, volcano ash, etc.
   • From near-real-time remote sensing imagery or model simulation

3. Dust as a climate indicator
   • Long-term satellite imagery, e.g. GOES, MODIS
   • Detect dust occurrence
   • Model training and mining process
   • Produce baseline for dust identification, trend, climate drivers, and predictability
Thank you!

myu7@gmu.edu
Scientific Question:
How to improve model efficiency?
Computational challenges

• Dust modeling is highly computing intensive
  • 72-hour, coarse-resolution (1/3 degree) for the U.S. Southwest using a single CPU: 4.5 hours
  • Resolution increasing to 1/9 degree: 5 days

• High performance computing:
  • 1/9 degree, 36 CPUs, <2hrs (Xie et al., 2010)
  • Decomposition and Parallelization
  • Communication

• Clustered allocation method: average 20% performance improvement (Huang et al. 2012)

Improve efficiency: Develop an optimized case-dependent subdomain collocation method!
Subdomain collocation algorithm

- K-Means and Kernighan-Lin combined Algorithm (K&K)

- Minimize total (or global) communication cost between computing nodes
- Balance workloads of computing nodes
- Balance communication among individual computing nodes

Applying K&K to dust modeling

Dividing a domain into finer scale subdomains cannot necessarily reduce execution time.
Allocate tasks on relatively low number of computing nodes, but also achieve high efficiency.
Performance Improvement Factor

\[ PIF = \frac{\Delta t}{t_{\text{default}}} \]

K\&K generates a regular subdomain division

Default allocation contains the largest possible communication
Indications of results

• Spatial collocation method
  • the granularity of subdomain
  • the optimized computing resource usage
  • achieve high efficiency

• Method not limited to dust simulation
  • Spatial optimization method:
    • Land use, regionalization, resource management
  • Other parallel computing tasks that require optimization on communication and computing costs
  • Extend to heterogeneous environments