Cloud-based modeling and big data approaches for ecosystem services: Moving from hype to substance

Ken Bagstad
WAVES Program, World Bank & U.S. Geological Survey
Today

• What is big data, and why does it matter now?
• Data-intensive science
• Semantics
• Why is cloud-based semantic modeling a challenge?
• Some proposed starting points for our community
Big data: Moving from hype to reality

Expectations

Overpromise
Reasonable expectations

Hype
Disillusionment

Ignorance

Education/expectation management

Experience & knowledge
What is big data?
Why are we here?

1. Ecosystem services going mainstream
2. Rise of computing power & computer science
3. Proliferation of new data sources
4. We’re moving quickly, but not quickly enough
1. Mainstreaming ecosystem services

http://www.wavespartnership.org
1. Mainstreaming ecosystem services

Additional countries active in natural capital accounting
1. Mainstreaming ecosystem services

Interest in: cross-jurisdictional management; evaluating tradeoffs; increasing responsivity to the public

Same, but short on resources & in-depth technical expertise
2. Rise of computing power & computer science

Moores law

CC BY-SA 3.0, shigeru23
2. Rise of computing power & computer science

CC BY-SA 3.0, Raul654
3. Proliferation of new data sources

“There’s no question we’re in a golden age for remote sensing.”

- Michael Freilich, NASA Earth Science Director


https://sentinel.esa.int/web/sentinel/home
3. Proliferation of new data sources

http://rmgsc.cr.usgs.gov/uas/
3. Proliferation of new data sources

Scientific Reports

Using social media to quantify nature-based tourism and recreation

Spencer A. Wood1,2, Anne D. Guerry1,3, Jessica M. Silver1,3 & Martin Lazo2

1Natural Capital Project, School of Environmental and Forest Sciences, University of Washington, Seattle, WA, USA, 2Natural Capital Project, Woods Institute for the Environment, Stanford University, Stanford, CA, USA.

Scientists have traditionally studied recreation in nature by conducting surveys at entrances to major attractions such as national parks. This method is expensive and provides limited spatial and temporal
4. We’re learning a tremendous amount about ecosystem services... but we’re not:

- Delivering in-time information to decision makers
- Driving down the time & cost of scientifically rigorous ES analysis fast enough
- Build off existing modeling studies & supporting meta analysis efficiently
Data-driven (i.e., inductive) science
Hey et al. 2009
1. Descriptive Science
2. Theoretical Science
3. Computational Science
Deductive science

Explanation why
Explanatory power ($r^2$)

**Good when you:**
Have strong understanding of how a system works
Want to test why a system works the way it does

**Pitfalls:**
Sloppy model construction
Not following best practices in modeling
4. Data-Intensive Science
Inductive science

Explanatory power ($r^2$)
Explanation why

Good when you:
- Want to explore for patterns
- Want to see if real-world patterns conform to theory
- Have incomplete datasets or high uncertainty

Pitfalls:
Putting too much faith into patterns found in the data that lack a reasonable theoretical foundation
Semantics in modeling
CAROLI LINNAEI, SPECI, DOCTORIS MEDICINÆ. SYSTEMA NATURÆ, SIVE REGNA TRIA NATURÆ SYSTEMATICE PROPOSITA PER CLASSES, ORDINES, GENERA, & SPECIES.

LUCDUNI BATAVORUM. AEC THEODORUM HAAK. MDCCXXXV.

EX TYPGRAPHIA JOANNIS WILHELMII OÙ GROOT.
Classifying land cover: A semantic jungle

<table>
<thead>
<tr>
<th>Data source</th>
<th>Spatial extent</th>
<th>Resolution</th>
<th># classes</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>SERVIR</td>
<td>Rwanda</td>
<td>30 m</td>
<td>13</td>
<td>1990, 2000, 2010</td>
</tr>
<tr>
<td>SERVIR</td>
<td>Tanzania</td>
<td>30 m</td>
<td>19</td>
<td>2000, 2010</td>
</tr>
<tr>
<td>ESA-CCI</td>
<td>Global</td>
<td>300 m</td>
<td>35</td>
<td>2000, 2005, 2010</td>
</tr>
<tr>
<td>GlobeLand</td>
<td>Global</td>
<td>30 m</td>
<td>10</td>
<td>2010</td>
</tr>
</tbody>
</table>

Manually harmonizing multiple datasets takes time & introduces subjectivity.
Semantics: The way out of the data & models jungle
A semantic approach to ecosystem services
(Villa et al. 2014)
Key design principles for semantic modeling

1. Keep semantics as modular as possible

2. Reuse existing expert knowledge wherever possible

3. Formalize semantics in *ontologies**

*Concepts & their interrelationships precisely defined
Anything we can observe (with data) has a subject

• Countable, physical, recognizable object

• Examples:
  1. A mountain
  2. A population of humans
  3. A population of trees (i.e., a forest)
  4. A river
Typical data describe a subject’s specific *quality*

- Described by an *observer type* (e.g., measurement, count, percentage, proportion...)
- Examples:
  1. The elevation of a mountain (measurement)
  2. Per capita income of a group of humans (value)
  3. Percent tree canopy cover (percent)
  4. A river’s stream order (ranking)
Over time, subjects may experience *processes*

- Examples:
  1. Erosion of a mountainside
  2. Migration of a human population
  3. Tree growth in a forest
  4. Streamflow in a river
A single, time-limited process is an event

• Examples:
  1. A snowfall event on a mountain
  2. The birth of a new human in the population
  3. The death of a tree in the forest
  4. A flood event on a river
*Relationships* connect two subjects

- Examples:
  - Structural connection – *Parenthood* connects *parents* to *children*
  - Functional connection – *Ecosystems* providing a particular benefit to *human beneficiaries*
- Very important for agent-based models
Observables can also have one or more traits

- “Adjectives” that add descriptive power to further modify a concept
- Add flexibility without adding more complexity to the ontologies
- Four types:

1. Attributes (Temporal, frequency, min/max/mean, etc.)
2. Identities (Authoritative species or chemical names)
3. Strata (Soil, atmosphere, ocean, forest)
4. Orderings (High-Moderate-Low)
A language to define data for environmental modeling

- Subjects
- Qualities
- Processes
- Events
- Relationships
- Traits
Why hasn’t data-intensive, cloud-based ecosystem service modeling taken off on its own?
Why is cloud-based modeling hard?
Why is cloud-based modeling hard?

Private goods

Public goods

Advertising

Environment

Health care
Why is cloud-based modeling hard?

I don’t understand the value of big data!

My data are more valuable to me if I keep others from using them!
Why is cloud-based modeling hard?
Some proposals on where to start
Standard data & modeling flow

• First time running an environmental modelers:
  - Download & preprocess input data
  - Parameterize & run models
  - Calibration, validation, sensitivity analysis
  - Policy analysis; disseminate results to stakeholders & scientific community
  - Archive data and models using metadata standards

• Second, third, fourth time:

**Same thing!**

At the end of the project, budgets are tight, and people want to get the paper/report out without worrying about proper archiving
Semantic data & modeling flow

• First time modelers:

  Download & preprocess input data
  
  Semantically annotate data & models
  
  Calibration, validation, sensitivity analysis
  
  Parameterize & run models
  
  Policy analysis; disseminate results to stakeholders & scientific community

• Archival work is **up front**, after that it’s usable by anyone

• Second, third, fourth time modelers:

  Parameterize & run models
  
  Calibration, validation, sensitivity analysis
  
  Policy analysis; disseminate results to stakeholders & scientific community

  Parameterize & run models
Cloud-based data and models: Toward context-aware modeling

- Many global & national datasets served by Web Coverage Service (WCS)/Web Feature Service (WFS)
- Data export via CKAN
Automating data & model assembly: Opens door for machine reasoning, pattern recognition

Set context: Runoff, Big Island, Hawaii

Query available models

Found 2 models: Wet & dry tropics appropriate

IF annual precipitation < 2,500 mm RUN MODEL A, OTHERWISE RUN MODEL B

Find & query annual precipitation layer

Identify data needs to run each model (DEM, precipitation, soils, vegetation, etc.)

Identified 3 digital elevation models (DEMs)

Select appropriate datasets to populate models; run models

Integrated modeling results & provide provenance diagram

Dem 1: Global DEM, 90 m resolution, reliability score = 70

Dem 2: State DEM, 10 m resolution, reliability score = 85

Dem 3: Study area DEM, 5 m resolution, reliability score = 90

Data integration

Run Model A

Run Model B

Raw precipitation data

Reclassified precipitation data
Making smarter choices about data needs for natural capital accounting

Consider needed spatiotemporal resolution and data quality

Metrics obtainable by satellite & radar

Input or calibration data for ecosystem service models (natural science & socioeconomic data)

Metrics obtainable in situ (incl. citizen science)

Adequate spatiotemporal resolution for ES assessment?
At a lunchtime plenary session at ACES 2012, Fort Lauderdale

“Who else is working on big data and semantic modeling in ecosystem services?”
Big data: Moving from hype to reality

Diagram:
- Overpromise
- Hype
- Ignorance
- Disillusionment
- Reasonable expectations
- Education/expectation management
## Getting from here to there

<table>
<thead>
<tr>
<th>Individual scientists &amp; practitioners can:</th>
<th>Attributes of a cloud-based, big data ecosystem service modeling system</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Understand what big data is and isn’t</td>
<td>1. Be modular &amp; independently expandible</td>
</tr>
<tr>
<td>2. Get comfortable with use of data-intensive science/inductive modeling when it’s appropriate</td>
<td>2. Be semantically consistent</td>
</tr>
<tr>
<td>3. Collaborate on and use semantics</td>
<td>3. Support multiple modeling paradigms</td>
</tr>
<tr>
<td>4. Share, serve, annotate data to facilitate cloud-based semantic modeling</td>
<td>4. Support context-aware modeling (AI &amp; machine learning), i.e., be data-driven</td>
</tr>
<tr>
<td>5. Fund computer science-based working on ecosystem services</td>
<td>5. Balance speed and accuracy (quick assessments/novice users), and support for high-level modeling (scientists/advanced users)</td>
</tr>
</tbody>
</table>
Competing 10-year visions

<table>
<thead>
<tr>
<th></th>
<th>Status quo</th>
<th>Semantic modeling*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time for natural capital assessment</td>
<td>6-12 months</td>
<td>1-2 months</td>
</tr>
<tr>
<td>Cost of natural capital assessment</td>
<td>$100K-1M</td>
<td>$20-50K</td>
</tr>
</tbody>
</table>

*Requires $10M investment in semantic standards & web services

• If ecosystem services become a national & international standard requested in everyday government decision making, can we afford the status quo?

• If not, how do we finance the needed investment in semantic modeling?
Thanks!

kbagstad@usgs.gov