Big Data for Spatial Analysis

D G Rossiter

Cornell University, Soil & Crop Sciences Section

Nanjing Normal University, Geographic Sciences Department
南京师范大学地理学学院

May 4, 2020

1. What is “big data”?
2. Examples
3. Portals
4. Uses of Big Data
5. Applications
6. How to deal with big (spatio-temporal) data?

What is “big data”?

Often discussed in terms of **volume** (size of data), **velocity** (frequency of data), and **variety** (diversity of data types):

volume data sets that are **too large** to be handled by common processing methods

- e.g., exceed the storage capacity of main memory or even secondary memory

velocity hyper-temporal data sets, or high-bandwidth streams of data

- e.g., social media activity, real-time locations

variety data sets that have too much **complexity** to be handled by common processing methods

- e.g., observations with 100’s to 1000’s of attributes of variable or even unknown data quality

What is different about big data?

- Too large to fully understand summaries or identification of unusual cases (“outliers”)
- Too complex to fully understand or control modelling
- Generally need machine-learning methods to analyze (e.g., random forests, neural nets)

Orders of magnitude

"Kilobytes $2^{10} \approx 10^3$ were stored on floppy disks. Megabytes $2^{20} \approx 10^6$ were stored on hard disks. Terabytes $2^{30} \approx 10^9$ were stored in disk arrays. Petabytes $2^{40} \approx 10^{12}$ are stored in the cloud.

"As we moved along that progression, we went from the *folder* analogy to the *file cabinet* analogy to the *library* analogy to – well, at petabytes we ran out of organizational analogies."

– Anderson, C. (2008, June 23). *The end of theory: the data deluge makes the scientific method obsolete*. **Wired**. https://www.wired.com/2008/06/pb-theory/ [1]

Problems using Big Data

1. **capturing** massive amounts of obserations
   - e.g., real-time sensors, satellite images, monitoring instruments
2. **storage** and data transfer from storage to computation
3. **search**, i.e., query to find/subset/summarize
4. **processing**, i.e., computation (CPU, memory, parallelization)
5. **analysis**: methods and understanding the results
6. **sharing**, information policies, e.g., privacy
   - different parts of the database may have different policies
7. **visualization**
   - summarizing with appropriate graphic design

Sources of big data

- sensor networks with fine temporal resolution
- mobile devices (e.g, geo-located phones)
- remote sensors; ever-increasing spatio-temporal resolution
- digital laboratory instruments
- point-of-sales or service (e.g., pharmacies, retail stores)
- user contributions (social media, citizen science)

Examples of large datasets

- NOAA $\approx 50 \cdot 10^9$ multivariate observations of oceans (as of 2018)[1]
  - temperature, salinity, oxygen, nitrates, phosphates and silicates at the particular **location** and **depth** collected at a particular **time**, so 4D
- 23andMe[2] DNA analysis of $5 \cdot 10^6$ individuals
- Twitter has about $500 \cdot 10^6$ tweets per day[3]
- MasterCard processed $74 \cdot 10^9$ transactions per year in 2012[4]

---

[1] https://www.nodc.noaa.gov/OC5/woa18/

[2] https://www.23andme.com/en-int/dna-ancestry/, 16-April-2019

[3] http://www.internetlivestats.com/twitter-statistics/

[4] http://blog.unibulmerchantservices.com/how-mastercard-processes-74b-transactions-a-year/

Citizen science

- eBird[5] (Cornell Ornithology): $590 \cdot 10^6$ observations (as of end 2018)
  - (semi-)automated quality control[6]

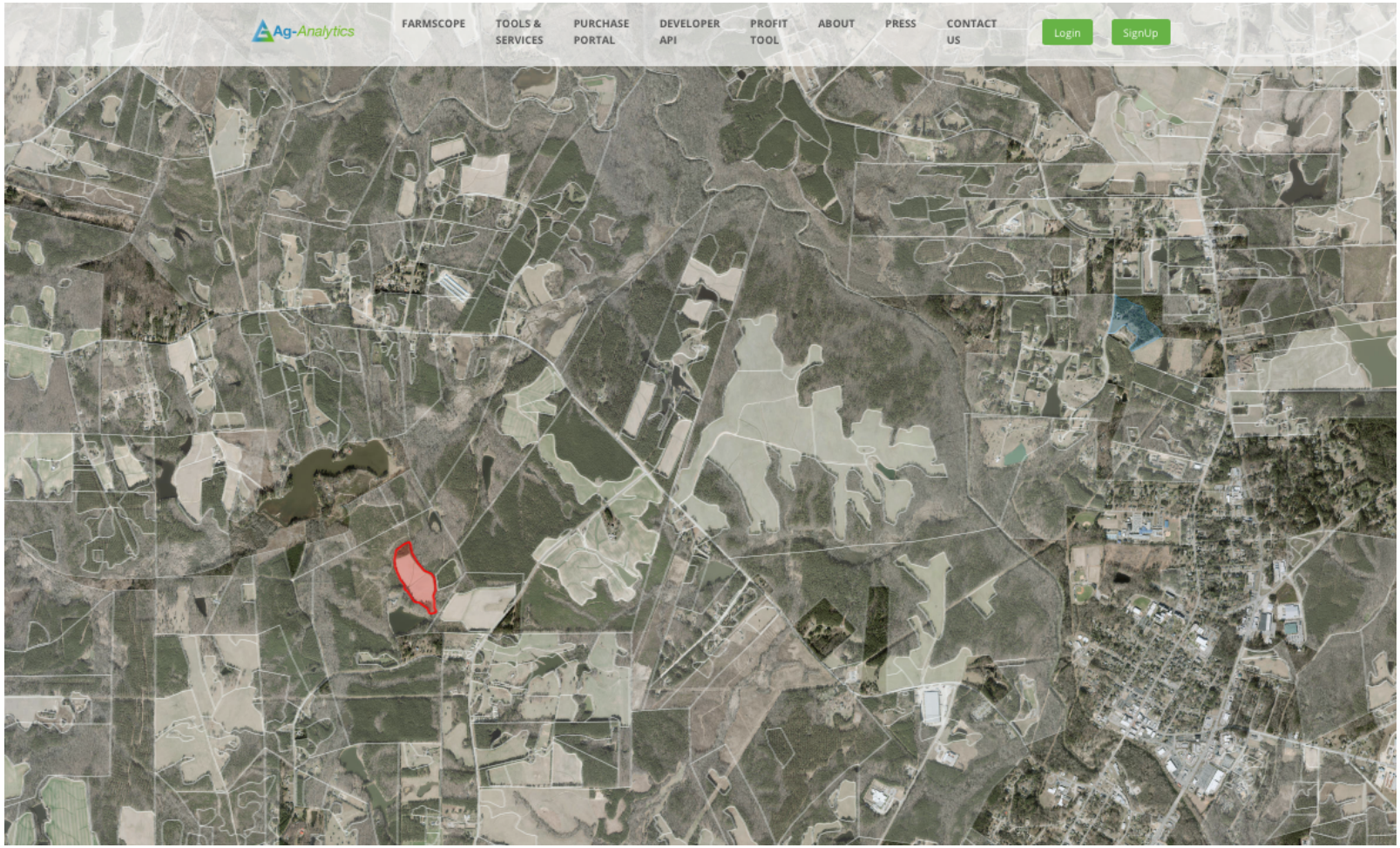
[5] http://www.ebird.org
[6] https://support.ebird.org/en/support/solutions/articles/48000795278

Agriculture

- Ag-Analytics[7]
  - see: “about”, “farmscope” maps
  - zoom in, click on individual fields: insurance, revenue, yield forecast
  - combines multiple open layers, with own analytics
- Gro Intelligence[8]
  - “leading the modern agricultural revolution using data and technology, driven by advances in parallel processing, remote sensing, machine learning, and AI”

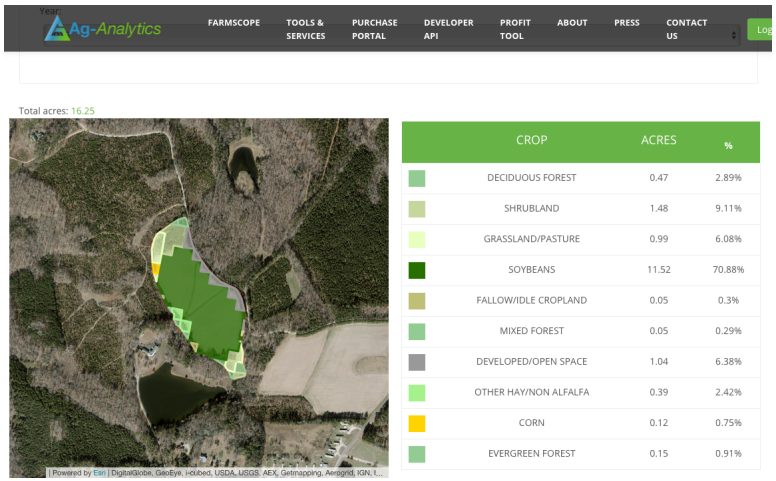
[7] https://ag-analytics.org/FarmScope
[8] https://www.gro-intelligence.com

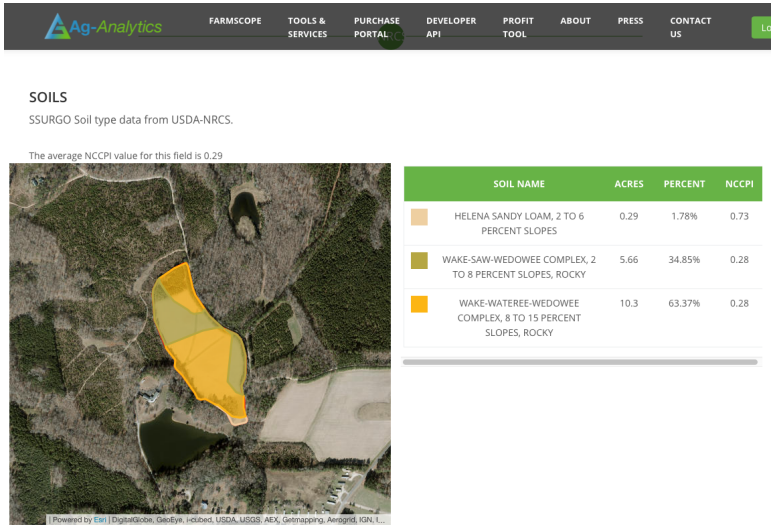
Ag-Analytics – location

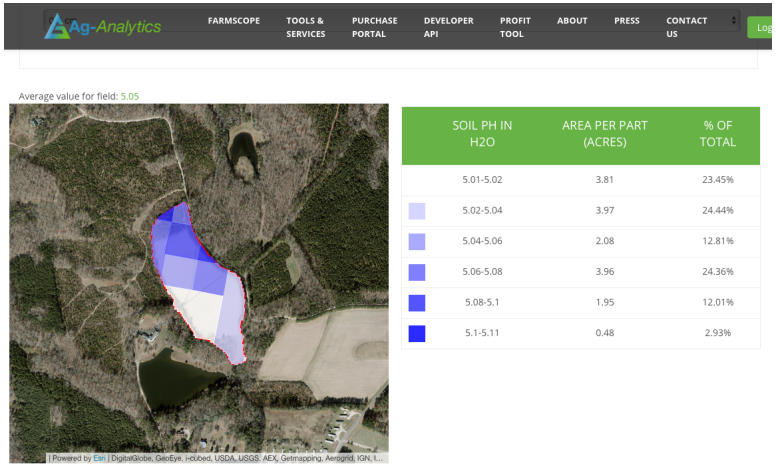


near Franklinton, North Carolina
field at (36.1068N, -78.3530E)

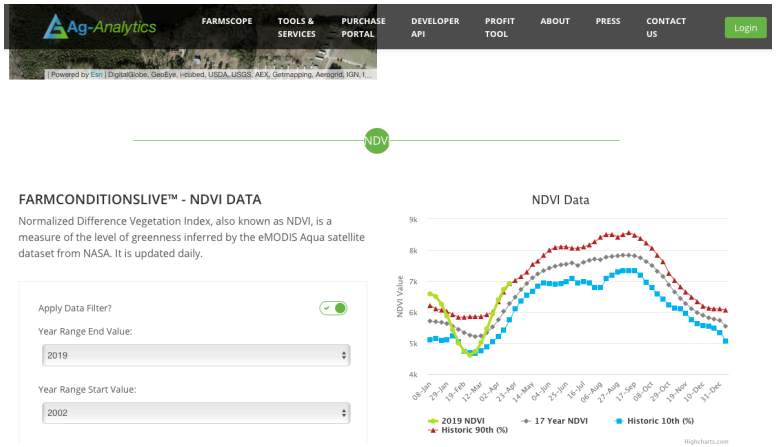
Ag-Analytics – field static information

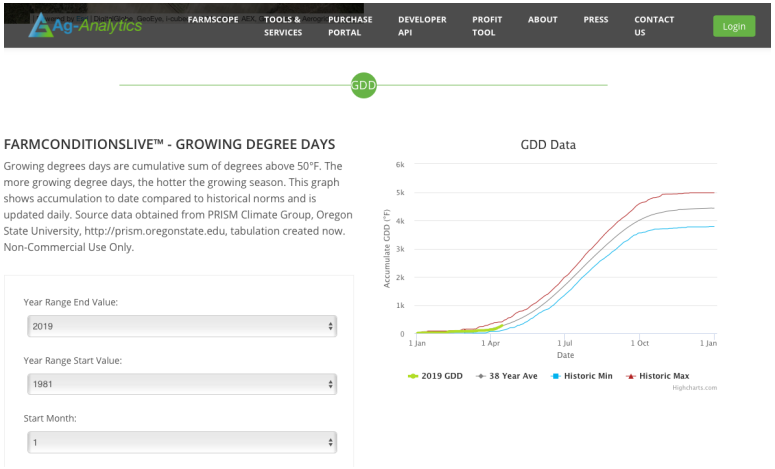


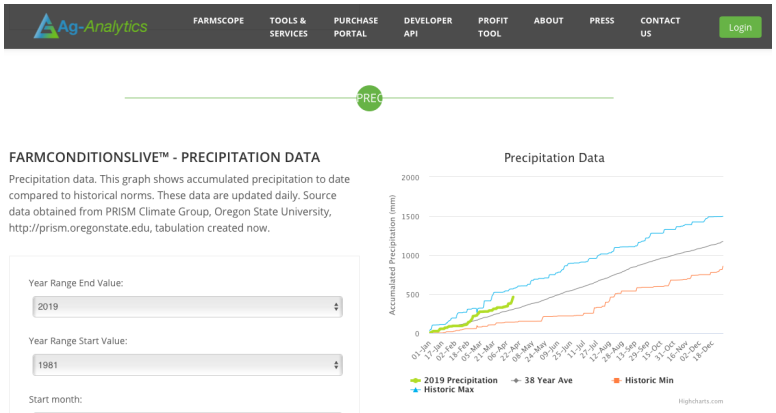




Ag-Analytics – field dynamic information







Gro Intelligence

Browse data

Search data

Visualize data from our collection of over 6.8 million data series and growing.

Geography
Land, Water and Land ›

Infrastructure
Production Infrastructure, Storage, Telecommunications ›

Investment, Lending and Transfers
Assets, Investment, Lending ›

Macroeconomic Indicators
Industry Indicators, Inflation, Labor Statistics ›

Prices
Indices, Market Prices, Producer Prices ›

Supply
Market Supply, Post-Harvest Processing, Production ›

Trade
Trade Flow Matrix, Trade

Weather and Climate
Data Infrastructure

Vast Range of Sources

Meaningful Geospatial Data

Analytical Tools

Visualizations

Soils – SoilGrids

- Reference: [5]
- 2017 version: global predictions for numeric soil properties: OC, bulk density, Cation Exchange Capacity (CEC), pH, soil texture fractions and coarse fragments at seven standard depths (0, 5, 15, 30, 60, 100 and 200 cm; total 280 raster layers
- based on $\approx$ 150k profiles, 158 gridded covariates
- New version May 2020[9]

[9] https://soilgrids.isric.org

SoilGrids processing flow

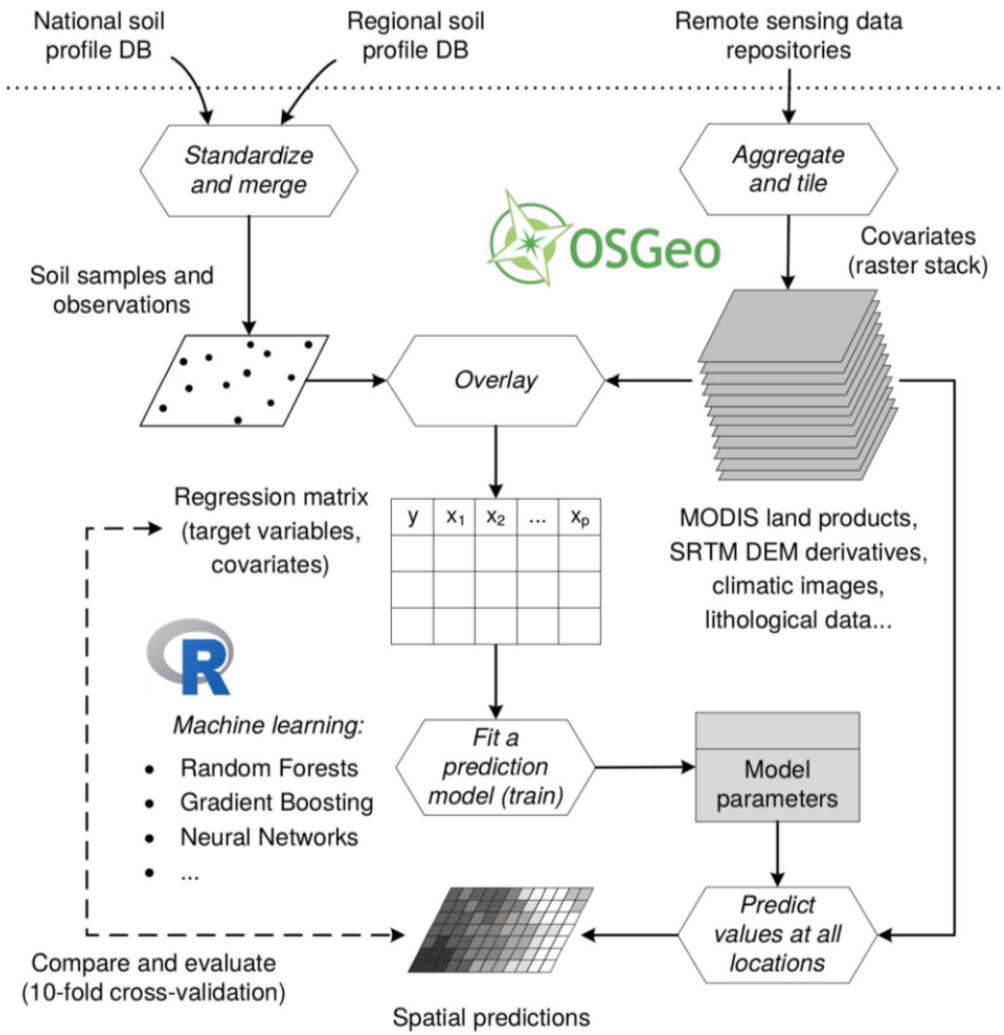


Fig 5. The (data-driven) statistical framework used for generating SoilGrids. SoilGrids are primarily based on publicly released soil profile compilations, NASA's MODIS and SRTM data products and Open Source software compiled with the ATLAS library: R (including contributed packages), and Open Source Geospatial Foundation (OSGeo) supported software tools.

doi:10.1371/journal.pone.0169748.g005

SoilGrids processing results

PLOS ONE | SoilGrids250m: Global gridded soil information

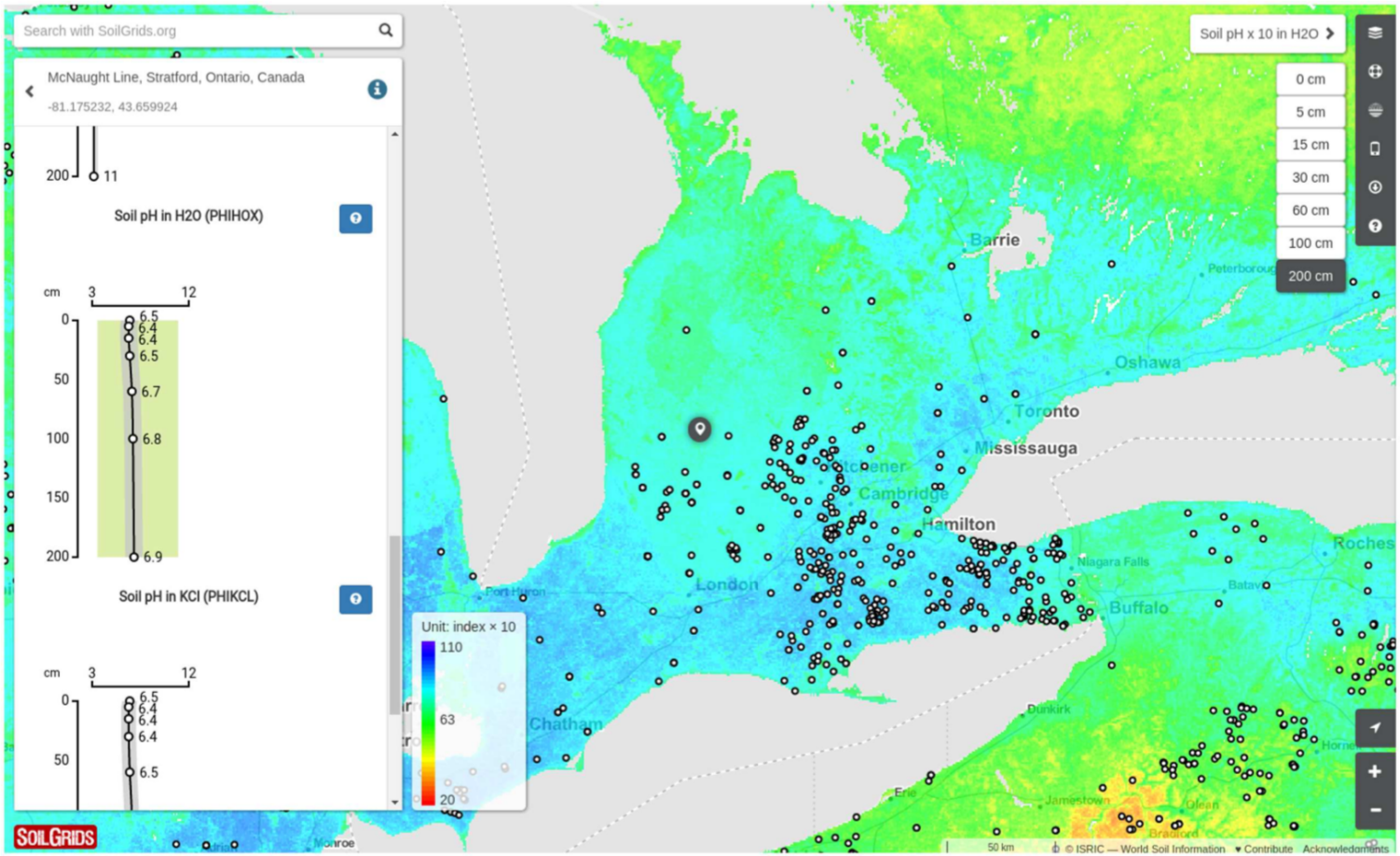


Fig 14. Basic design and functionality of SoilGrids.org: Soil web-mapping browser that provides interactive viewing of 3D soil layers. Reference administrative data, basic functionality and output data license of SoilGrids.org are primarily based on OpenStreetMap.

doi:10.1371/journal.pone.0169748.g014

Soils – POLARIS

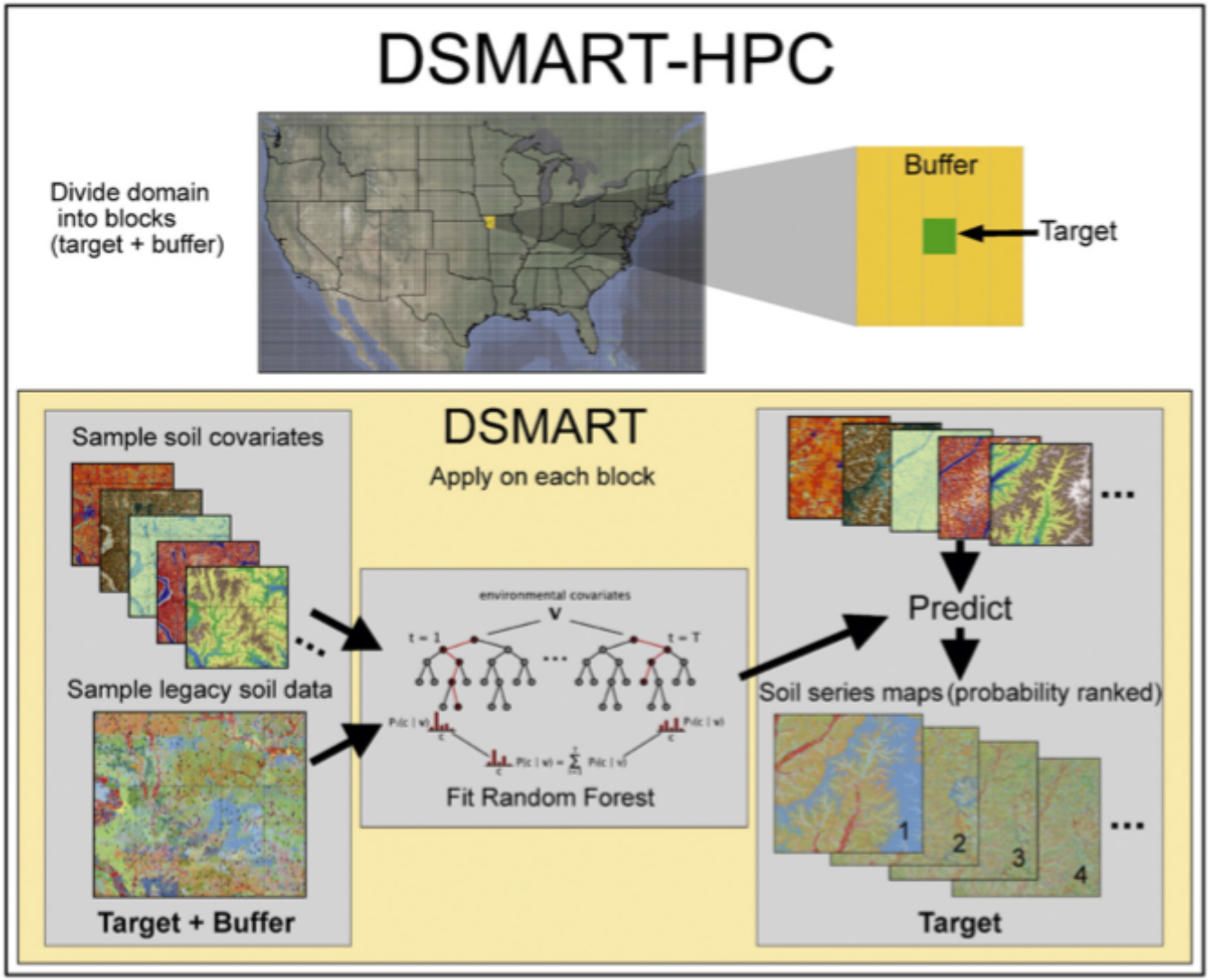
POLARIS [3]

- Purpose: map **probability** of **soil series** at **every grid cell** in the lower 48 (USA) States; total $\approx 1.25 \cdot 10^9$ grid cells
- algorithm: DSMART (Disaggregation and Harmonization of Soil Map Units Through Resampled Classification Trees)
- supercomputer "Blue Waters"; 12 474 nodes 30 x 30 km, with 60 km buffer to ensure continuity
- 30 m horizontal spatial resolution, so 1000 x 1000 = 10^6 per 1° tile[10]
- required 450 000 core-hours = 5 wall-clock hours
- "This is negligible computer time at current HPC facilities that can handle 10 million ($\approx$ 1100 years) core-hour tasks."

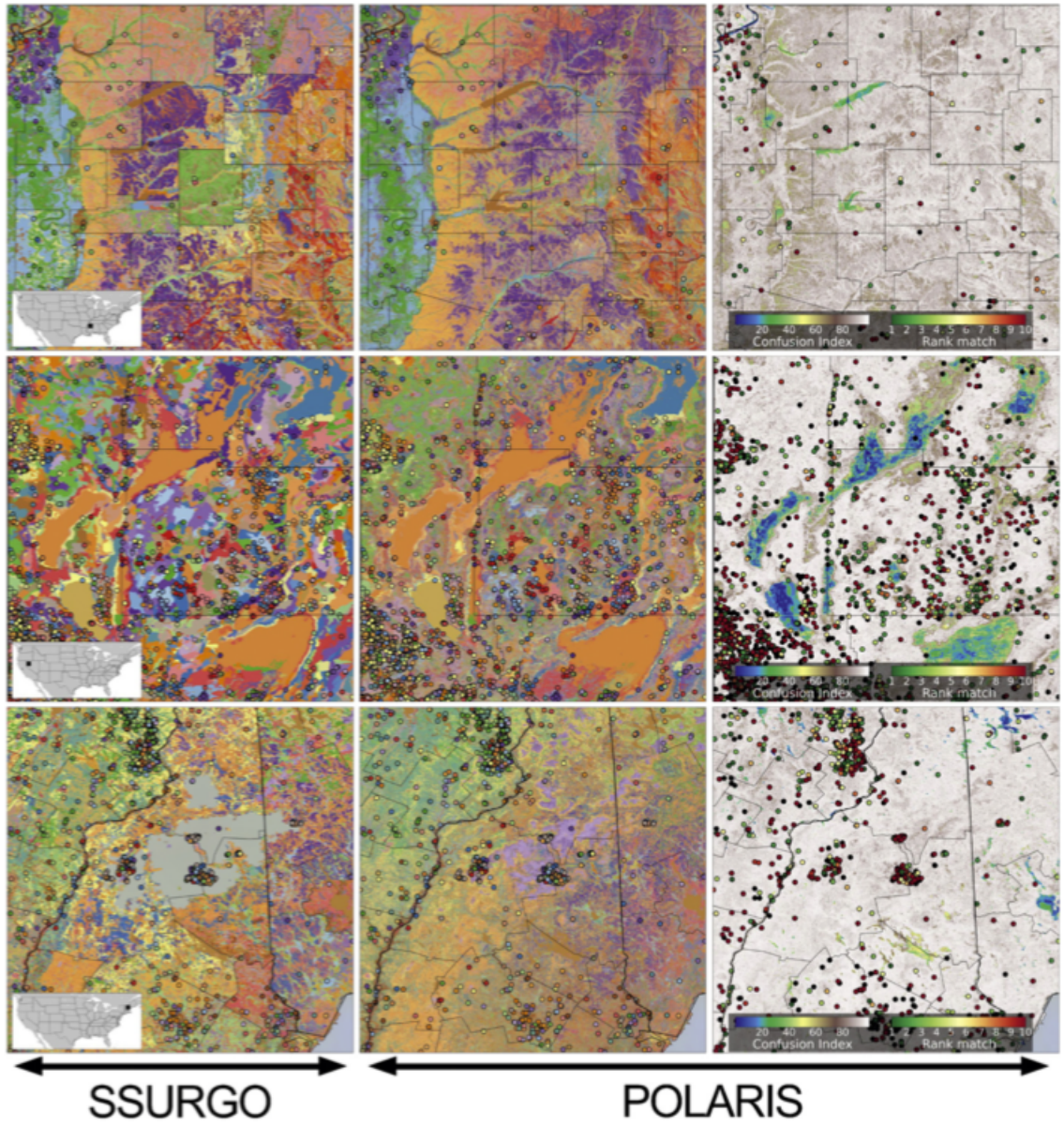
[10] http://stream.princeton.edu/POLARIS/

POLARIS methods

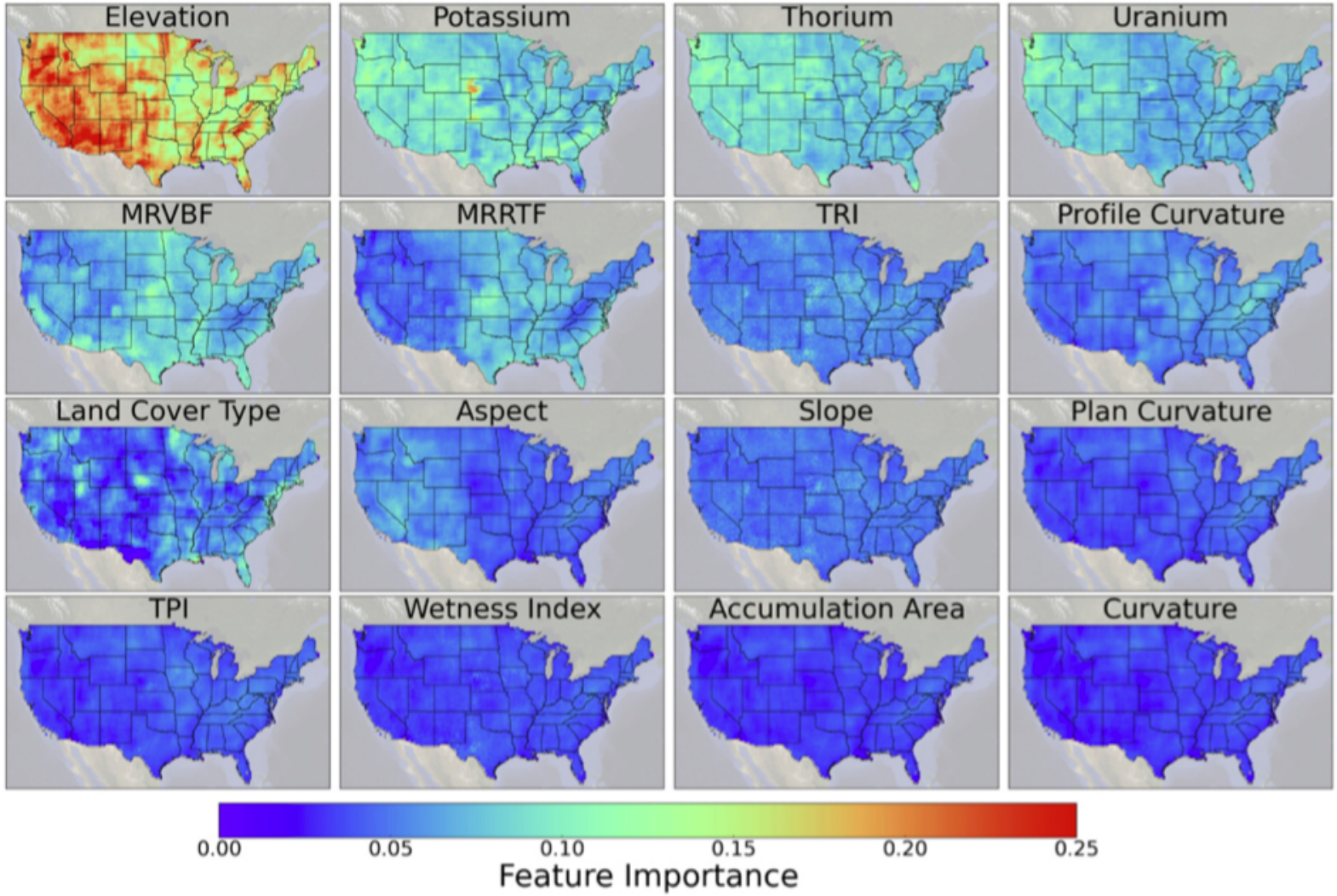
N.W. Chaney et al. / Geoderma 274 (2016) 54–67



POLARIS results



POLARIS covariate importance



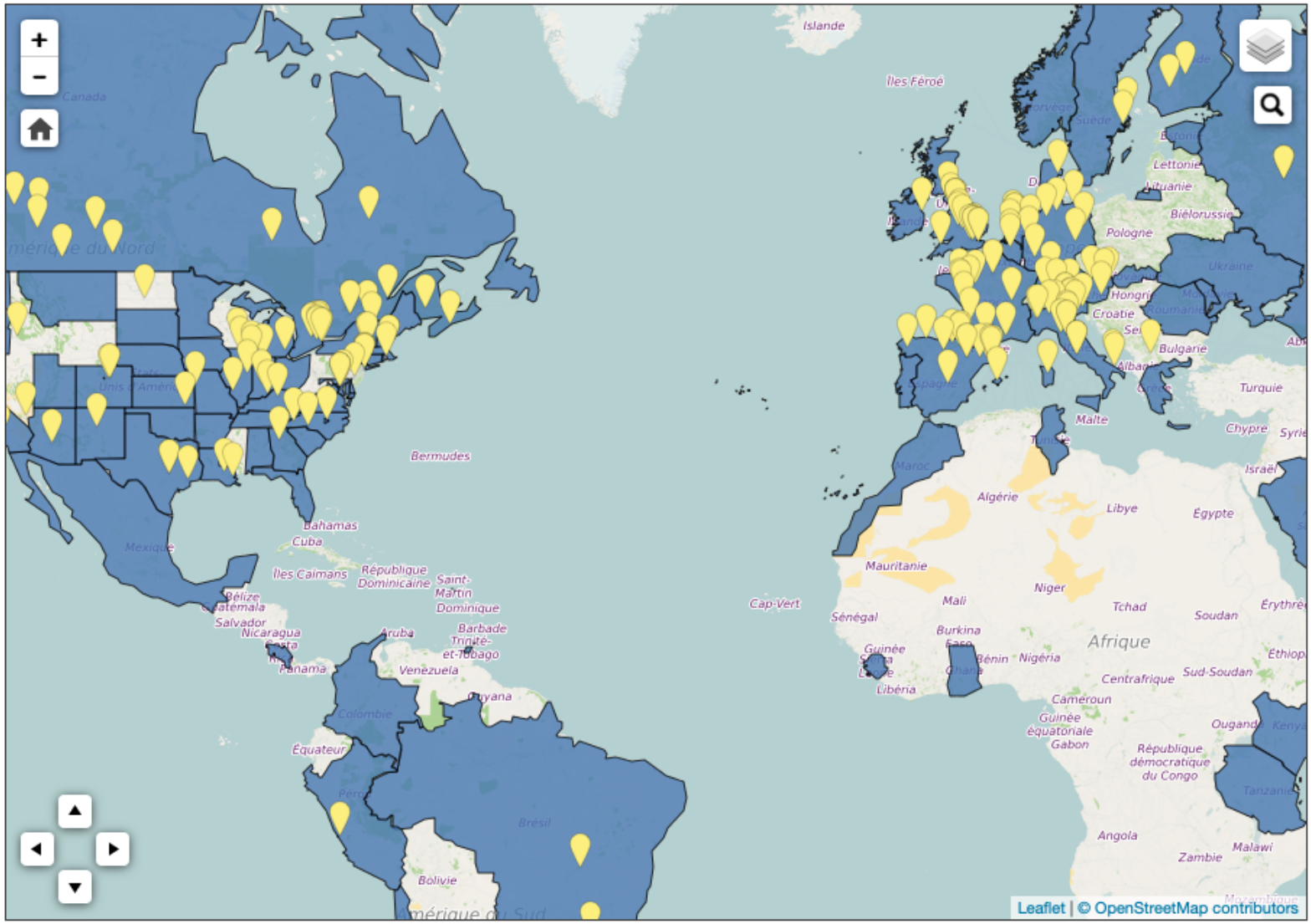
Portals and data brokers

- Collect and catalog many sources of geographic data
- example: World Food Programme (WFP) Geonode[11]
- example: Soil Geographic Databases[12]
- example: OpenGovernment[13]
- problem: **searching** for relevant (to the user) information
  - WFP Geonode: search by region, type of information, keyword, date, extent, file type

[11] https://geonode.wfp.org
[12] https://www.isric.org/explore/soil-geographic-databases
[13] https://www.data.gov/open-gov/

Portal: OpenGovernment

Map representation of Open Data Sites

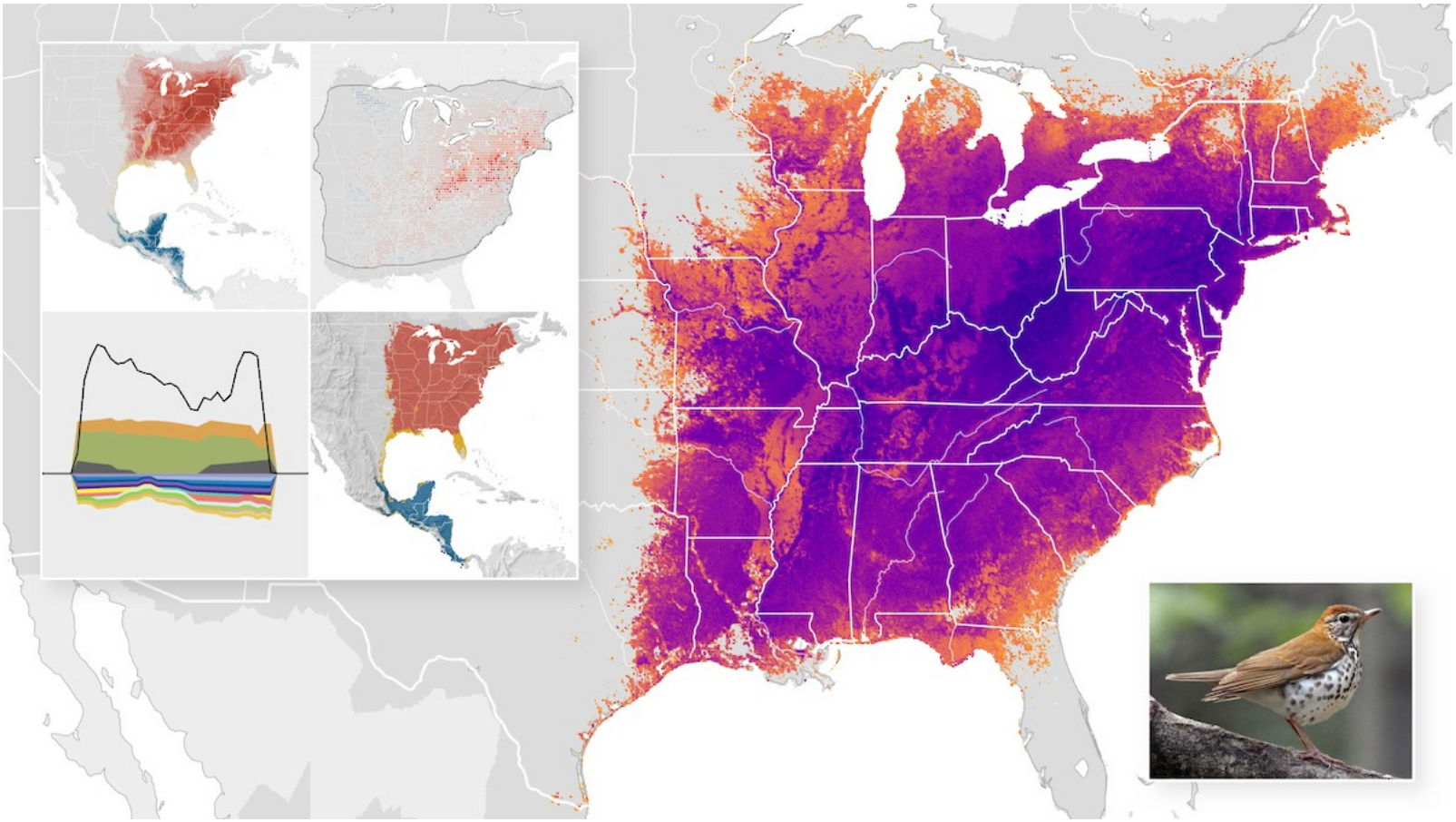


Uses of Big Data

- Scientific research into processes
  - e.g., eBird Science applications[14]:
  - "connect[s] birdwatchers around the world in a way that informs research and conservation . . . 2019 featured the first annual update of eBird Status and Trends, which now provides status and distribution information for 302 species ... "[15]
- Mapping
  - e.g. SoilGrids, POLARIS
- Prediction, decision support
  - e.g., Ag-Analytics, Gro Intelligence
- Visualization, hypothesis formation

[14] https://ebird.org/science
[15] https://ebird.org/news/ebird-2019-year-in-review

Analysis: eBird analytics



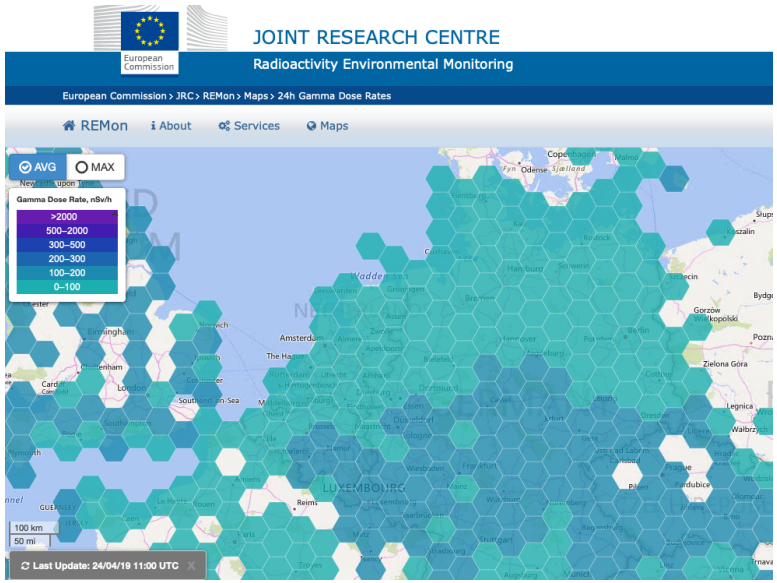
Status & trends for wood thrush (*Hylocichla mustelina*)

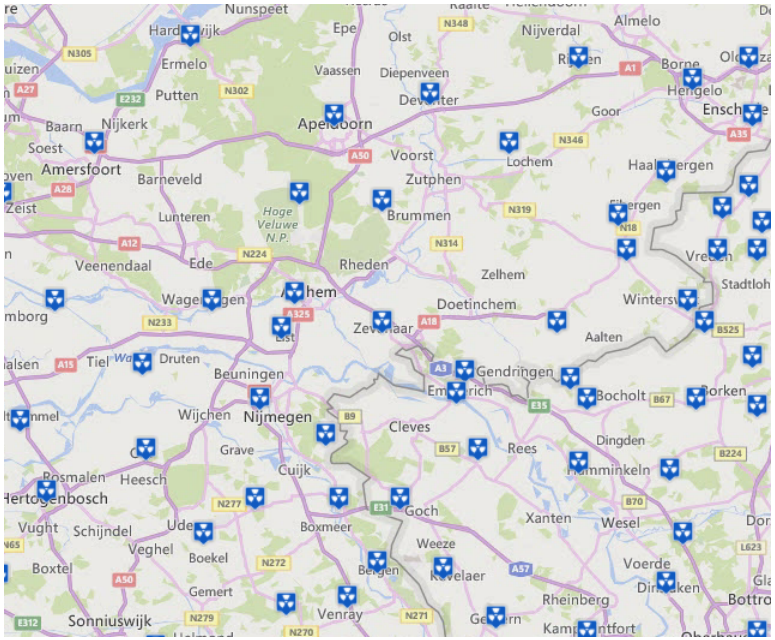
Prediction, (near) real-time

- European Radioactivity Environmental Monitoring[16]
  - "Gamma dose rate averages and maxima for the last 24 hours in almost real time"
  - European Radiological Data Exchange Platform (EURDEP): a **network** for the **exchange** of radiological monitoring data between most European countries
  - Large network of sensors, automatic reporting and summarizing
- Methods: see [6] "**Real-time automatic interpolation** of ambient gamma dose rates from the Dutch radioactivity monitoring **network**"

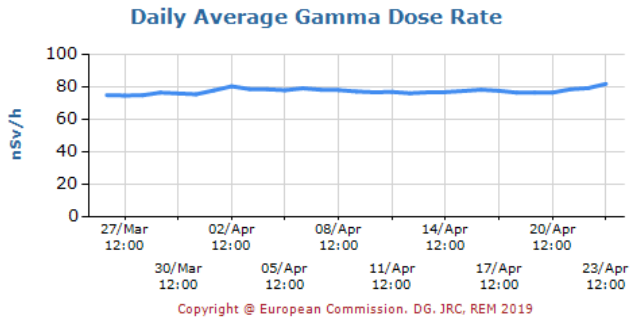
[16] https://remap.jrc.ec.europa.eu

Radiation monitoring





ARNHEM



Some applications

- General discussions: [2, 4]
- Ecology: [7]
- Geographic sociology: [8]
- Epidemiology: [10, 12]
- Agroecosystems: [9]
- Agricultural entomology: [11]
- Radiation monitoring: [6]

How to deal with big (spatio-temporal) data?

- **automated** methods for data capture, data screening (quality control)
- **robust** statistical methods, not sensitive to outliers, not dependent on manual selection of model form
- **massive** computing resources
- **collaborative** science – combine different disciplinary knowledge; also requires experts in inter-, multi-disciplinary collaboration

Computing platforms for big data

Massive datasets can not be handled on many personal/departmental computers.
So **cloud computing** must be used.

- Example: Google Earth Engine[17]
  - "A **planetary-scale** platform for Earth science data & **analysis** – Powered by Google's **cloud infrastructure**"
  - "hosts satellite imagery and stores it in a **public data archive** that includes **historical earth images** going back more than **forty years** . . . made available for global-scale **data mining**.
  - Aimed at consistent Earth-wide analyses, but can be used regionally or locally

[17] https://earthengine.google.com/

Use of GE Engine

- Computation is all done **remotely** (parallel processing)
  - local computer only for coding/viewing
- Accessible via an Application Programming Interface (**API**)
  - Javascript, Python, **R** with rgee package[18]
- Built-in code editor
- Direct access to the datasets
- **Image processing**, **Geometry** algorithms
- **Machine-learning** algorithms: un/supervised classification

[18] https://csaybar.github.io/rgee/

GE Engine datasets

Many datasets are also too large for local storage, so they are also in the cloud and used as needed; Google Earth Engine includes many datasets[19], including:

- Imagery (Landsat, Sentinel, MODIS . . .)
- Atmospheric conditions (can help correct other products)
- Weather
- Geophysical: terrain (e.g., SRTM), elevation
- Nighlights
- Administrative
- Interpreted: land cover, land use, cropland (e.g., USDA NASS; Global Food Security)

[19] https://developers.google.com/earth-engine/datasets/catalog

Conclusion

Ever-increasing amounts of data, ever-increasing computer power, allow integrating many “big” data sources to deal with “big”, complex, inter-disciplinary problems.

References I

[1] Chris Anderson. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, Jun 2008. ISSN 1059-1028. URL https://www.wired.com/2008/06/pb-theory/.

[2] Roger Bivand and Konstantin Krivoruchko. Big data sampling and spatial analysis: "which of the two ladles, of fig-wood or gold, is appropriate to the soup and the pot?". *Statistics & Probability Letters*, 136:87–91, May 2018. doi: 10.1016/j.spl.2018.02.012.

[3] Nathaniel W. Chaney, Eric F. Wood, Alexander B. McBratney, Jonathan W. Hempel, Travis W. Nauman, Colby W. Brungard, and Nathan P. Odgers. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma*, 274:54–67, Jul 2016. doi: 10.1016/j.geoderma.2016.03.025.

[4] Hamid Ekbia, Michael Mattioli, Inna Kouper, G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R. Sugimoto. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8):1523–1545, 2015. doi: 10.1002/asi.23294.

References II

[5] Tomislav Hengl, Jorge Mendes de Jesus, Gerard B. M. Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N. Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, and et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2):e0169748, Feb 2017. doi: 10.1371/journal.pone.0169748.

[6] Paul H. Hiemstra, Edzer J. Pebesma, Chris J.W. Twenhöfel, and Gerard B.M. Heuvelink. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Computers & Geosciences*, 35(8):1711–1721, Aug 2009. doi: 10.1016/j.cageo.2008.10.011.

[7] S. L. LaDeau, B. A. Han, E. J. Rosi-Marshall, and K. C. Weathers. The next decade of big data in ecosystem science. *Ecosystems*, 20(2):274–283, Mar 2017. doi: 10.1007/s10021-016-0075-y.

[8] Jiwei Li, Qingqing Ye, Xuankai Deng, Yaolin Liu, and Yanfang Liu. Spatial-temporal analysis on Spring Festival travel rush in China based on multisource big data. *Sustainability*, 8(11):UNSP 1184, Nov 2016. doi: 10.3390/su8111184.

[9] M. Susan Moran, Philip Heilman, Debra P. C. Peters, and Chandra Holifield Collins. Agroecosystem research with big data and a modified scientific method using machine learning concepts. *Ecosphere*, 7(10):e01493, 2016. doi: 10.1002/ecs2.1493.

References III

[10] Dirk U. Pfeiffer and Kim B. Stevens. Spatial and temporal epidemiological analysis in the big data era. *Preventive Veterinary Medicine*, 122(1–2):213–220, Nov 2015. doi: 10.1016/j.prevetmed.2015.05.012.

[11] Jay A. Rosenheim and Claudio Gratton. Ecoinformatics (big data) for agricultural entomology: pitfalls, progress, and promise. *Annual Review of Entomology*, 62(1):399–417, Jan 2017. doi: 10.1146/annurev-ento-031616-035444.

[12] Kim B. Stevens and Dirk U. Pfeiffer. Spatial modelling of disease using data- and knowledge-driven approaches. *Spatial and Spatio-temporal Epidemiology*, 2(3):125–133, Sep 2011. doi: 10.1016/j.sste.2011.07.007.