A multi-variate examination of the cause(s) of increasing induced seismicity events using geoscience computing & big data techniques

Kelly Rose, Jennifer Bauer, D.V. "Vic" Baker, Devin Justman, Roy Miller III, Aditya Tiwari - National Energy Technology Laboratory, U.S. Department of Energy

Office of Research and Development

NETL

Abstract/Overview

The "Induced Seismicity, Geoscience Computing & Big Data" project is part of DOE's SubTER Crosscut initiative. Our approach is to utilize NETL's Energy Data eXchange (EDX) resources and data mining techniques to assess the use of "big data" for understanding the causes and future probability of induced seismic events and their relationship to a range of possible natural and anthropogenic triggers. Results to date include gathering and integration of key datasets needed to support the goals of the study, evaluation of enhancements and changes needed for the EDX system to support more robust and efficient data mining, integration and rapid analytical needs in support of the project. These goals are as follows:

- Develop tools and approaches to manage multiple heterogeneous datasets
- Develop a probabilistic approach that can be utilized to assess potential for induced seismicity impacts through big data analyses
- Develop stochastic approaches to reduce uncertainty and constrain subsurface trends
- Improve joint analysis of multiple datasets, focus on advancing "Big Data" mining and integration techniques to improve knowledge and reduce uncertainty about subsurface systems through advanced geoscience computing
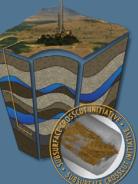
Data Collection

Climate Records
Census data and infrastructure
Topography
Aquifer data
Hydrology, geology, and fault data
695,050 Oil, gas, and groundwater wells
Hydrocarbon reservoir data
Over 17,000 seismic events

Data Volume

Over 154 data volumes with varying spatial and temporal extents and resolutions

Over 635 GB of Data
Percent of data volume by category*

- Infrastructure 0.1%
- Hydrocarbons 2%
- Geographic 8%
- Climate 9.5%
- Hydrologic 0.5%
- Seismic 29.5%
- Well Data 50.4%

*Plus an additional 607.5 GB of Well Log Data

Temporal Data Coverage

Partial or incomplete data

- Monthly & Annual Injection Well Volume
- Monthly & Annual Oil/Gas Well Production Volume
- Monthly Temperature
- Monthly Precipitation
- Earthquakes

1885 1895 1905 1915 1925 1935 1945 1955 1965 1975 1985 1995 2005 2015
Year

N — 90 45 0 90 Miles

Data Interrogation and Analysis

In year one of this project, the NETL team conducted preliminary spatio-temporal analyses to investigate key datasets and attributes affecting induced seismicity. Examples are displayed below.

Time Vs. Earthquake Depth

Depth (km)

Plot showing Oklahoma earthquake events by depth versus time. Earthquake events with unconstrained hypocenters highlighted in blue ellipse.

Time (YYYY)

Source: OGS (Oklahoma Geological Survey)

Spatio-temporal Analysis

Source: OGS

Faults
- 2010 - 3/23/2015
- 2000 - 2009
- 1990 - 1999
- 1980 - 1989
- 1970 - 1979
- 1960 - 1969
- 1950 - 1959

Earthquakes in Oklahoma per Year 1950-2015

Years	Total Earthquakes
2010-2015	13120
2000-2009	208
1990-1999	133
1980-1989	521
1970-1979	188
1960-1969	66
1950-1959	22
Total	**15040**

Ellipses show spatio-temporal trends in Oklahoma for earthquakes per decade from 1950–March 23, 2015. For each decade, ellipses were calculated based off the standard deviation of X and Y coordinates from the mean center of earthquakes. Graph above represents earthquake occurrences per decade.

0 25 50 100 150 200 Miles

Multi-Variate analysis

Ellipses show distributions for earthquakes (1950-2010, 2010-3/23/2015), faults, and surface water (rivers, lakes, and streams). The earthquake distribution from 1950-2010, has similar orientations as surface water and fault distributions, while the earthquake distribution from 2010-2015 does not. For each variable, ellipses were calculated using the standard deviation of X and Y coordinates from the mean center of each dataset.

Source: OGS, OKWRB (Oklahoma Water Resources Board)

Faults
- 2010 - 3/23/2015 EQ Dist.
- 1950-2010 EQ Dist.
- Surface Water Distribution
- Fault Distribution

0 25 50 100 150 200 Miles

Subsurface Analysis

A regional geologic framework of Oklahoma is being assembled using well log correlations. The project has a total of 534,965 oil and gas wells that have 237,758 well logs. This example shows a cross section from western to North central Oklahoma, labeled as B – B'.

SS Depth (ft)

TOPS AND MARKERS
- MISSISSIPPIAN_CHAT_RM RM
- WOODFORD_RM RM
- MISSISSIPPIAN_LIME_RM RM
- WOODFORD DJ
- CLEVELAND_SS_RM RM
- BIG_LIME_RM RM
- COTTAGE_GROVE_RM RM
- OSWEGO_RM RM
- SYLVAN DJ
- HUNTON DJ
- VIOLA DJ
- SIMPSON DJ
- ARBUCKLE DJ
- ARBUCKLE RM
- CHECKERBOARD RM
- BASEMENT DJ

Source: IHS Inc.

Spatial Statistics

Source: OGS

EQ Magnitude
- Coldspots (blue)
- Hotspots (red)

Hot spot analysis based off earthquake magnitude identifies areas of statistically significant clusters (95% confidence level) of higher magnitudes (hotspots in red) and lower magnitudes (coldspots in blue). Results show a clear separation between hotspots in North-central Oklahoma, and coldspots in mid/central and Southern Oklahoma.

0 25 50 100 150 200 Miles

Word cloud: Wells, Hydrocarbons, Geography, Seismic, Injection, Production, Climate, Geology, Reservoirs, Infrastructure

Next Steps

In the next 18 months, the project team will include additional partners, and focus on developing i) new geoscience data capabilities and ii) a probabilistic approach for assessing the likelihood of induced seismicity.

Geoscience computing for big data is emerging, for this project we seek to advance capabilities for data gathering, integration, and analysis by developing an EDX/ Velo/ Cluster Computing architecture for this project & SubTER needs more widely.

Insights from these analyses have highlighted key parameters influencing induced seismicity as well as identified gaps in big data analytical approaches that must be addressed to support more robust and advanced analysis of engineered geosystems. Moving forward, this project will focus on analyzing these data and filling analytical and technology gaps to ensure sustainable, environmentally responsible access to domestic resources. Tasks for the next phase include:

Big data is often defined by one or more of the following: **volume** of the data, **velocity** at which the information is provided, **variety** of data, and **variability** of data sources.

BigData — Sources, DataDiversity, ComputingCapabilities, Metadata, DataDriven, DataMining, MultiScaleIngest, Information, Content, Disseminate

Interaction Workflow Diagram — Internal NETL Network — EDX — Hadoop — Velo

Task 1 – Geoscience computing advances
(NETL, LANL, PNNL, ANL)

1.1 - Develop an EDX / Velo / Hadoop integration to support data gathering, mining, and analytical needs (NETL, PNNL)

1.2 - Evaluate tools for irregular data management (ANL, LANL, NETL, PNNL)

1.3 - Integrate irregular data management capabilities (NETL, LANL, ANL, PNNL)

Task 2, Development of probabilistic approaches for evaluating induced seismicity
(LANL, LLNL, NETL, OGS, SNL)

2.1 - Data processing, reduce uncertainty seismic events (OGS, LLNL, NETL)

2.2 - Finalize data gathering & integration (NETL, ANL, SNL)

2.3 - Probabilistic analyses and uncertainty sources (NETL, SNL, OSU, LANL)

2.4 - Develop beta version of stochastic approach (NETL, SNL, LANL)