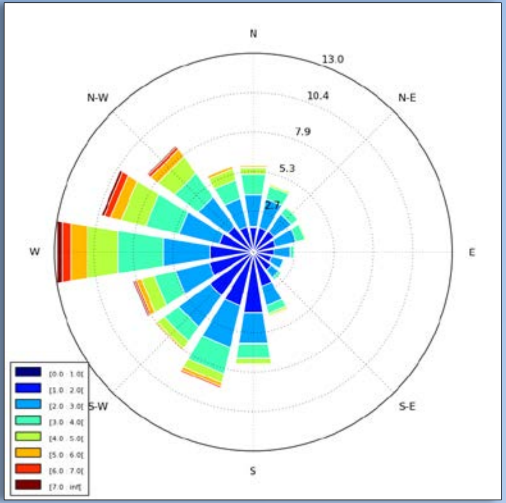
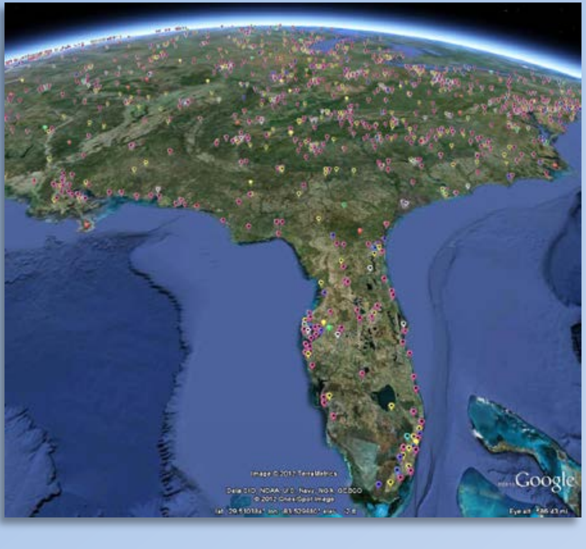
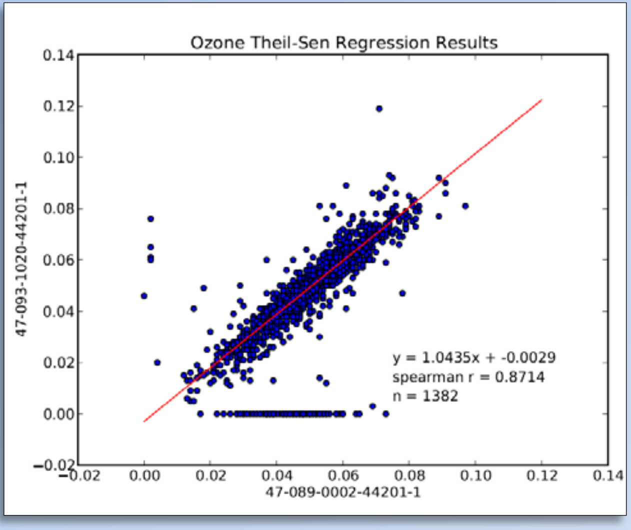
Air Quality Data Analysis Using Open Source Tools







Daniel Garver and Ryan Brown

US EPA Region 4

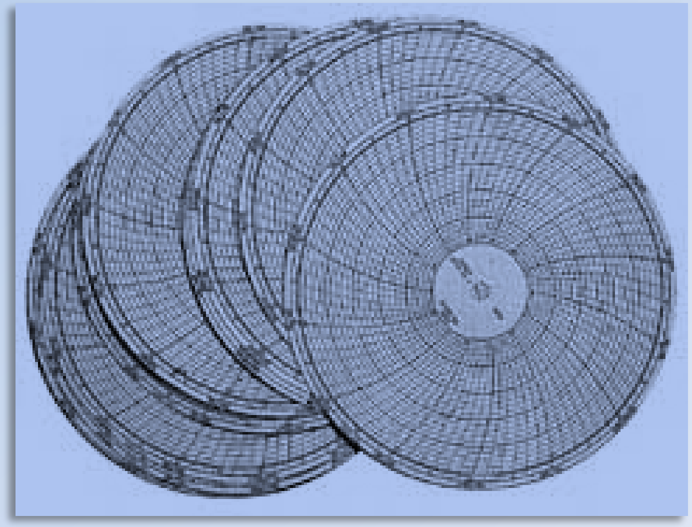
EPA United States Environmental Protection Agency

The Problem

- Do you ever want to do more than is possible through a standard AQS report? And have it automated?

- Do you wish you had additional software but don't have the resources?
  - Software like SAS, SigmaPlot, Matlab, etc. is useful for air quality data analysis.
  - For many agencies, however, these tools are cost prohibitive.

- Do you have routine reports or analyses that you currently generate manually and are resource/time intensive?



Outline

- About Python
- SQL databases and Python - SQLAlchemy
- Region 4 Examples:
  - – Ozone site correlations
  - – Air monitoring geodatabase
  - – Automated wind roses and pollution roses
- Other potential applications

```
"""icon.py - draw file ico
"""
import Image, ImageDraw, I
SHADOW = True

doc = Image.open('code.png')
draw = ImageDraw.Draw(doc)
py = open(__file__, 'r').rea
draw.text((32, 8), py)

if SHADOW:
  s = ImageOps.gray
  e = ImageEnhance
  s = e.enhance(0.
  s.paste(doc, (0,0
  doc= s

logo = Image.open('brand.
doc.paste(logo, (120,96), logo)
doc.save('py.ico')
```

What is Python?

EPA United States
Environmental Protection
Agency

python™

“Python is a programming language that lets you work more quickly and integrate your systems more effectively. You can learn to use Python and see almost immediate gains in productivity and lower maintenance costs.” - *Python.org*

- Extremely portable and compatible – the same code runs on Windows, Linux/Unix, Mac OS X and works with existing code in C, C++, Fortran, Java, R, and HTML
- Open Source License – free to use (unlimited number of users and licenses) and immediately available for many diverse applications
- Strong user base that continually develops and supports python – free add-ons available for database integration, graphing, math/statistics, web development (22,846 add-ons currently available and growing)

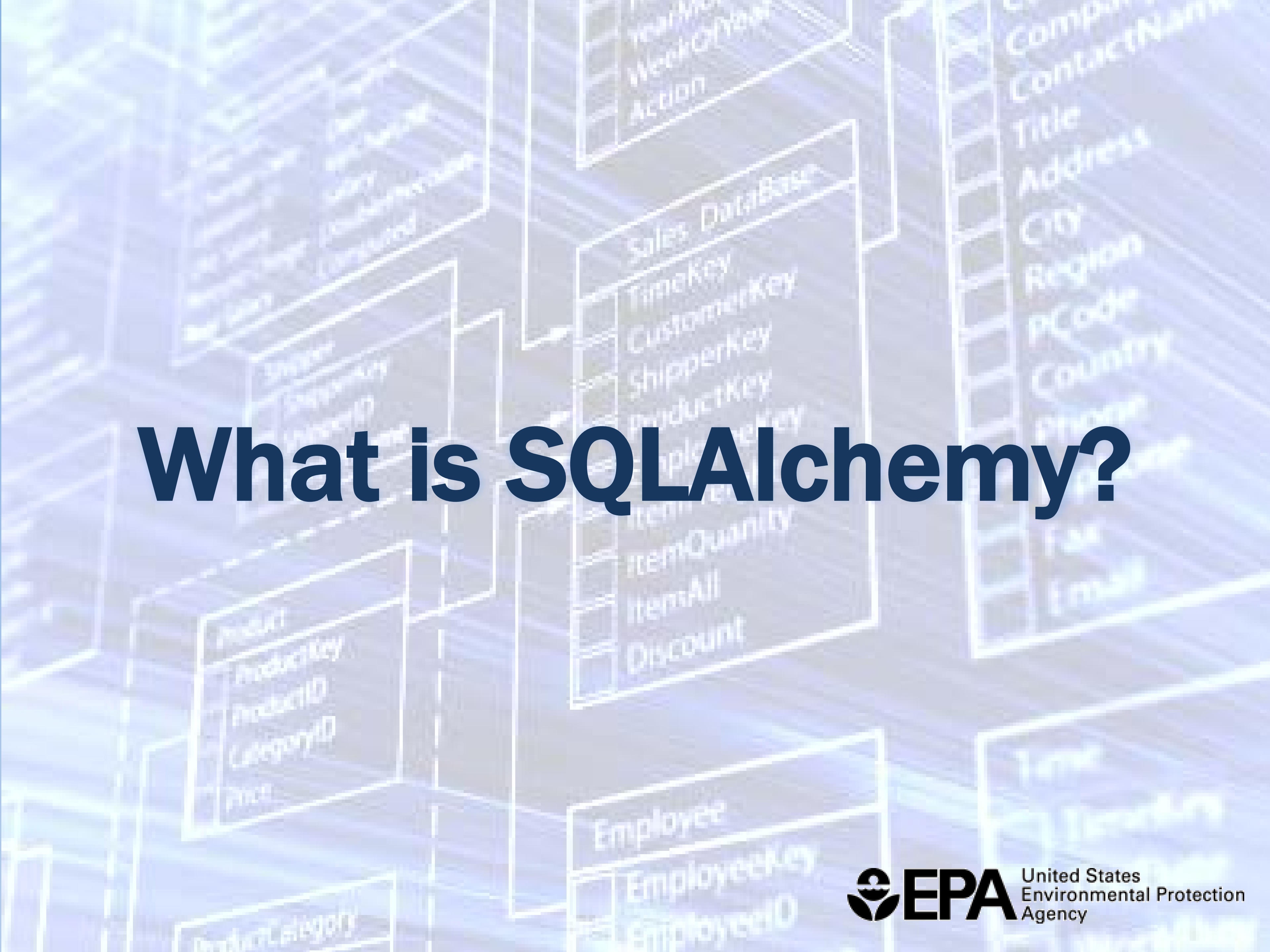
python™

Same functionality as widely available, costly, commercial software

- Free, open source add-ons similar to:
  - Matlab
  - SAS
  - PDF writing software
  - Graphing software
  - HTML Web Interface software
  - Database management/migration software

Using Python With Air Quality Data

- Air quality data is almost always stored in an SQL relational database
- Examples:
  - AQS, AQS Datamart (Oracle)
  - AIRNow (MySQL)
  - Proprietary air monitoring data management systems such as E-DAS or AirVision (Usually MS SQL Server or Oracle)
  - State or Local Agency Databases (various platforms)
- *The first step is to access the data using Python…*
  - *SQLAlchemy: a Python add-on*



What is SQLAlchemy?

Python and SQLAlchemy

- SQLAlchemy is a Python add-on that interacts with relational databases
- The same Python code can communicate with any type of SQL database (e.g. Oracle, MySQL, SQL Server, Access)
- Python code with SQLAlchemy is easier to read and write than raw SQL
- Your data can then be used for other purposes in Python

SQLAlchemy – An Example

This Python code…

```
query = session.query(dm.dim_monitor).join(dm.dim_facility)
```

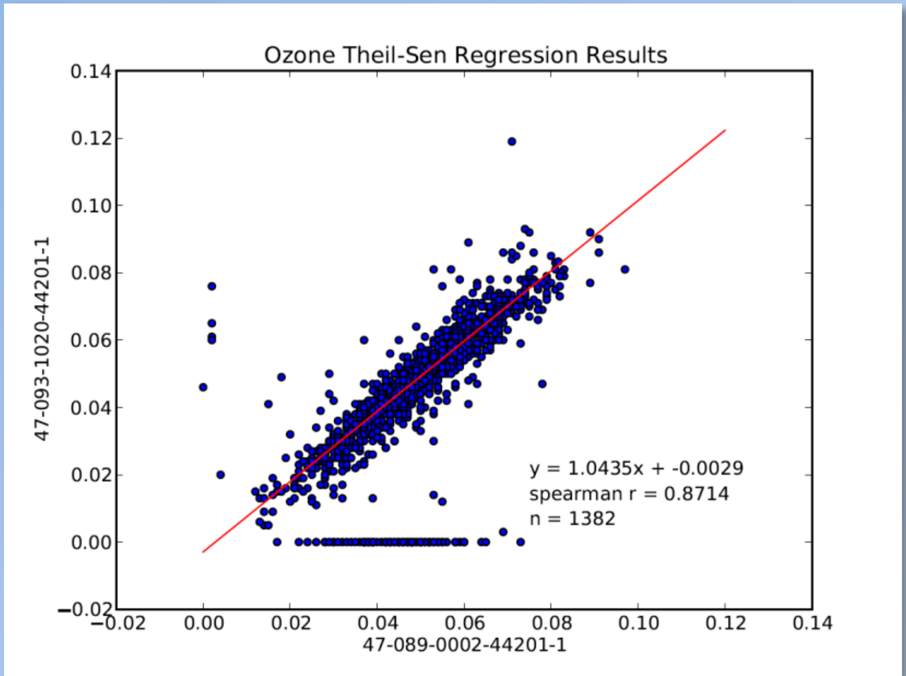
Generates this SQL to the AQS Data Mart:

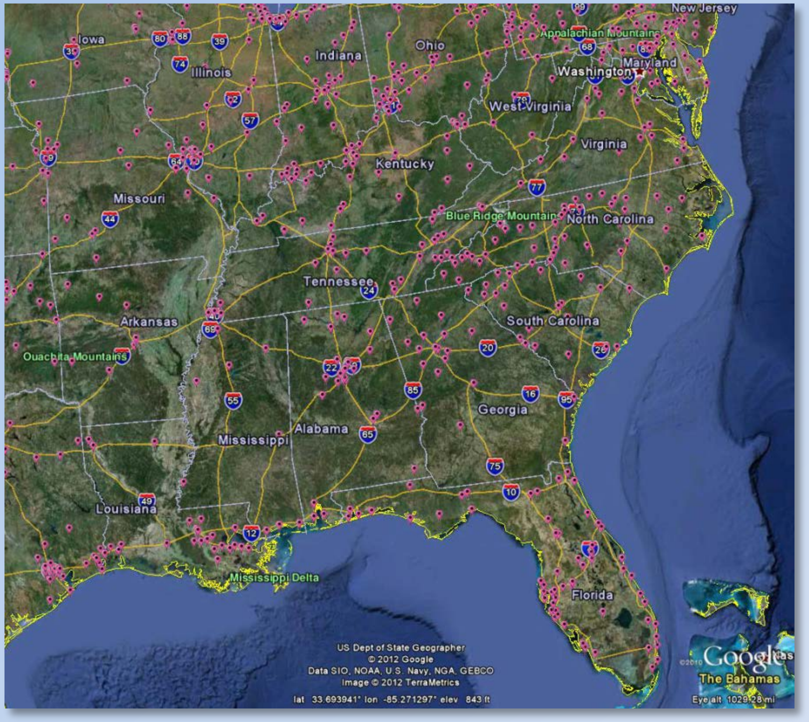
```
SELECT aqsmart.dim_monitor.dim_monitor_key AS aqsmart_dim_monitor_dim__1,
aqsmart.dim_monitor.dim_facility_key AS aqsmart_dim_monitor_dim__2,
aqsmart.dim_monitor.dim_substance_key AS aqsmart_dim_monitor_dim__3, aqsmart.dim_monitor.mo_id
AS aqsmart_dim_monitor_mo_id, aqsmart.dim_monitor.poc AS aqsmart_dim_monitor_poc,
aqsmart.dim_monitor.measurement_scale AS aqsmart_dim_monitor_meas_4,
aqsmart.dim_monitor.measurement_scale_definition AS aqsmart_dim_monitor_meas_5,
aqsmart.dim_monitor.project_type_code AS aqsmart_dim_monitor_proj_6,
aqsmart.dim_monitor.project_type AS aqsmart_dim_monitor_proj_7,
aqsmart.dim_monitor.dominant_source AS aqsmart_dim_monitor_domi_8,
aqsmart.dim_monitor.probe_location AS aqsmart_dim_monitor_prob_9,
aqsmart.dim_monitor.probe_height AS aqsmart_dim_monitor_prob_a,
aqsmart.dim_monitor.probe_horiz_distance AS aqsmart_dim_monitor_prob_b,
aqsmart.dim_monitor.probe_vert_distance AS aqsmart_dim_monitor_prob_c,
aqsmart.dim_monitor.sample_residence_time AS aqsmart_dim_monitor_samp_d,
aqsmart.dim_monitor.unrestr_air_flow_ind AS aqsmart_dim_monitor_unre_e,
aqsmart.dim_monitor.surrogate_ind AS aqsmart_dim_monitor_surr_f,
aqsmart.dim_monitor.collaborating_programs AS aqsmart_dim_monitor_coll_10,
aqsmart.dim_monitor.last_sampling_date AS aqsmart_dim_monitor_last_11,
aqsmart.dim_monitor.etl_last_load_process AS aqsmart_dim_monitor_etl__12,
aqsmart.dim_monitor.etl_last_load_date AS aqsmart_dim_monitor_etl__13
FROM aqsmart.dim_monitor JOIN aqsmart.dim_facility ON aqsmart.dim_facility.dim_facility_key =
aqsmart.dim_monitor.dim_facility_key
```

Once you have access to your air quality data in Python, you can:

- Perform statistical analysis
- Create graphs
- Export data for other applications (e.g. Excel)
- Display your data in a web interface
- Create PDF reports
- Load data into another database (e.g. for GIS)
- Perform geoprocessing
- Perform automated analysis for data validation or QA/QC

An AQS Example: Ozone Site Comparison Regressions (with Graphs!)





Can be run in minutes!   Continually useful!

Graphs!

United States Environmental Protection Agency

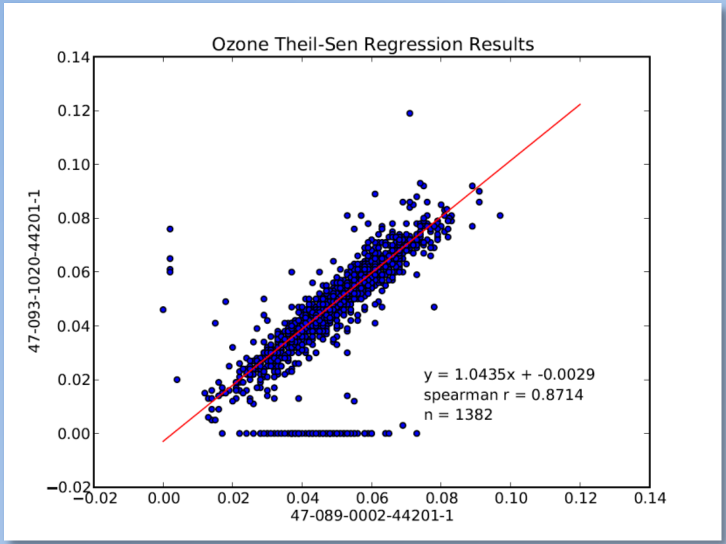
Ozone Site Comparison Regressions

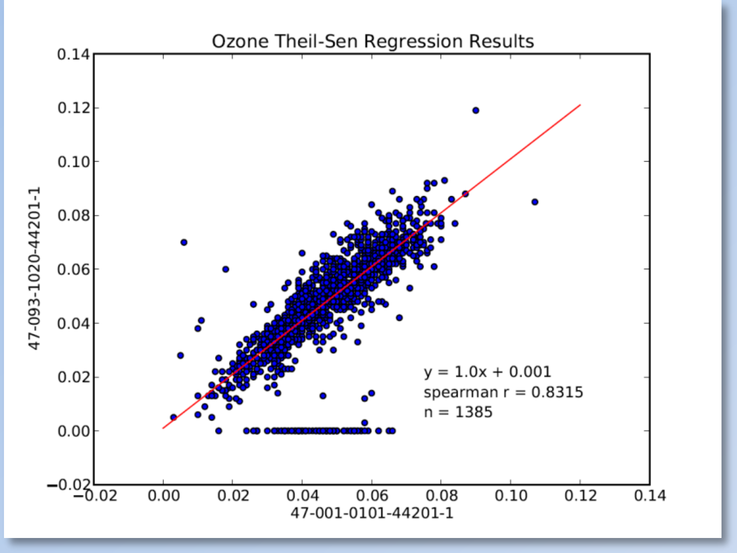
- Compares one site with all other sites in the CBSA based on daily maximum 8-hr average
- Generates nonparametric Theil-Sen Regression plots of each site pair
- Calculates correlation statistics
  - Spearman’s r, Pearson’s r, etc.
- Potential uses:
  - Evaluating network design (identifying redundant sites)
  - Modeling concentrations in the event of missing data

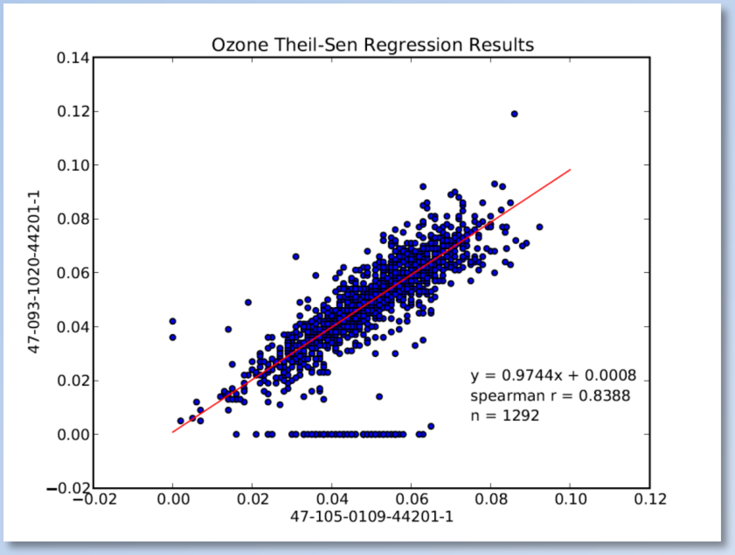
Ozone Site Comparison Matrix

Site A	Site B	n	Spearman r	Spearman p-val	Pearson r	Pearson p-val	Theil slope	Theil int
47-093-1020-44201-1	47-093-0021-44201-1	1444	0.951914952	0	0.943372755	0	0.94118	0.00185
47-063-0003-44201-1	47-093-0021-44201-1	292	0.941248585	9.46E-139	0.933572848	2.90E-131	0.91892	7.00E-05
47-105-0108-44201-1	47-093-0021-44201-1	93	0.912599594	4.10E-37	0.900784971	1.00E-34	0.75	0.01025
47-089-0002-44201-1	47-093-0021-44201-1	1417	0.877047948	0	0.733866562	6.64E-240	1	-0.001
47-105-0109-44201-1	47-093-0021-44201-1	1327	0.832461933	0	0.704831416	7.99E-200	0.88889	0.00411
47-001-0101-44201-1	47-093-0021-44201-1	1420	0.825591772	0	0.693255698	5.94E-204	0.92453	0.00421
47-009-0101-44201-1	47-093-0021-44201-1	1482	0.747143127	7.83E-265	0.641101197	2.45E-172	0.91667	-0.00271
47-155-0101-44201-1	47-093-0021-44201-1	1484	0.726548204	8.57E-244	0.619745174	3.62E-158	1.03571	-0.00966
47-155-0102-44201-1	47-093-0021-44201-1	1102	0.655271896	3.59E-136	0.474699528	5.18E-63	0.94118	-0.00524
47-009-0102-44201-1	47-093-0021-44201-1	1425	0.637935642	1.17E-163	0.535401114	1.65E-106	0.85714	0.00771

Regression Plots:
Selected Southeast Sites







Example:
An Air Monitoring Geodatabase

- Custom AQS Data Mart Queries
- New database design is created in geodatabase
- Data loaded into a geodatabase
- Python scripting in GIS can automatically create custom layers and feature classes

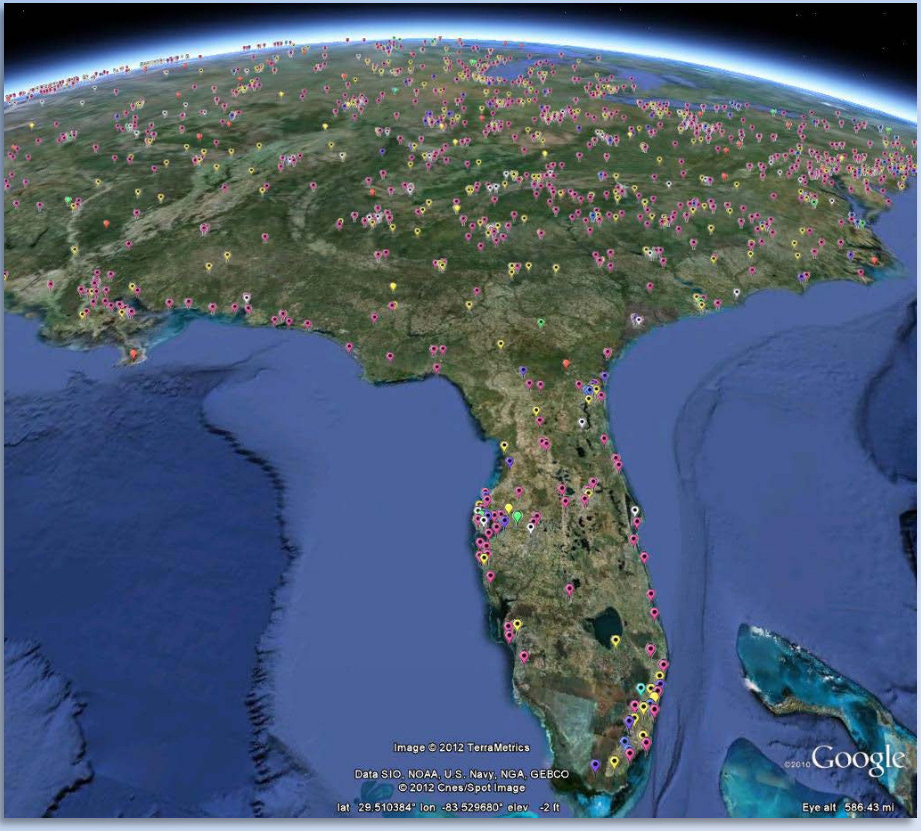
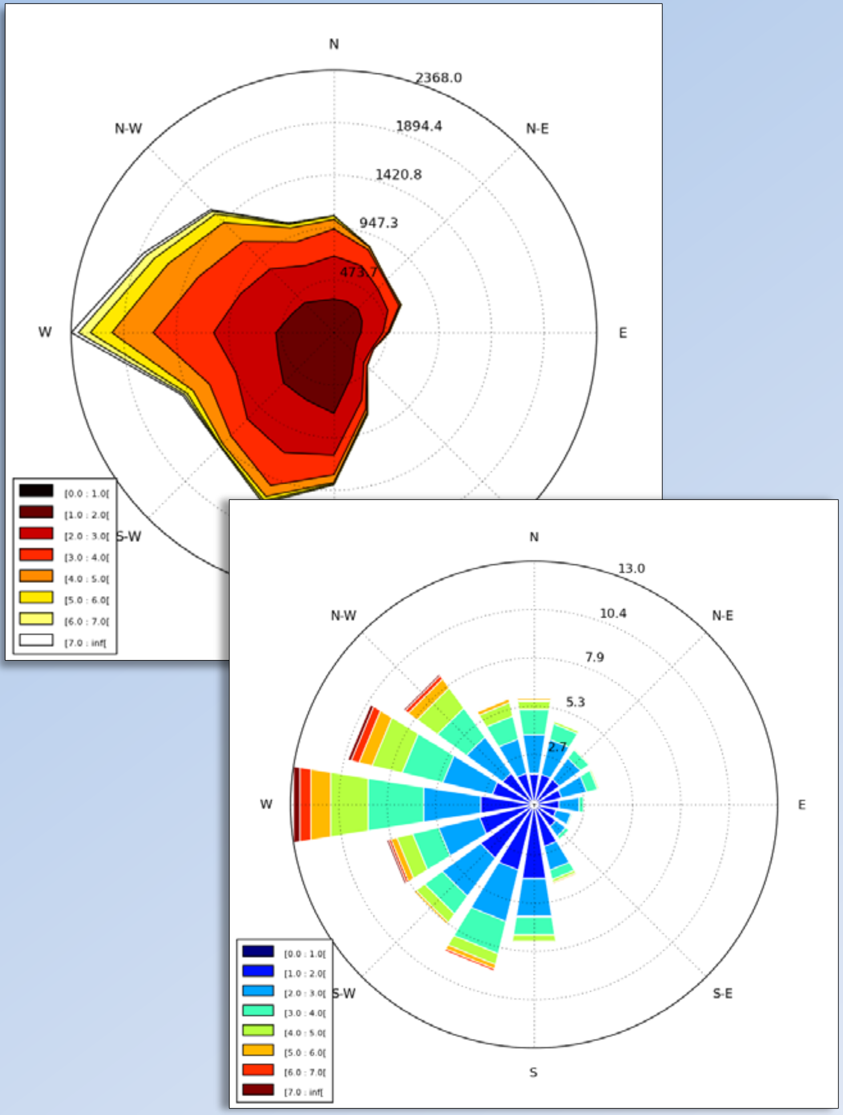


Image © 2012 TerraMetrics
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2012 Cnes/Spot Image
lat 29.510384° lon -83.529680° elev -2 ft
©2010 Google
Eye alt 586.43 mi

EPA United States Environmental Protection Agency

Example:
Air Pollution Wind Roses

- Customized AQS Data Mart query pulls:
  - Concentration data over a certain value (e.g. SO_2 > 75 ppb)
  - Corresponding wind speed and direction for those hours with high concentration
- Another Python add-on is used to generate a wind rose or pollution rose of the hours with high concentration



Other possibilities for air quality data

- Object wrappers/framework for retrieving data from specific databases (e.g. AQS Datamart)
  - Could design complex queries with less coding
  - Speed up development time for specific analysis, graphing, web development functionality
- Python powered web interface
  - Data retrieval, analysis/graphics, and web connectivity can be coded all together

Summary

- Agencies rely on air quality data analysis for planning and decision making
- State, Local, and Federal Budgets are being cut
- Data analysis automation can provide real value
  - Time savings
  - Better, more reliable analysis
- Python is a free tool that can help achieve this goal:
  - Easy to learn and read
  - Ideal for scientists and engineers (part time programmers)

Additional Resources

- Daniel Garver
  - garver.daniel@epa.gov
  - 404-562-9839
- Ryan Brown
  - brown.ryan@epa.gov
  - 404-562-9147

- Python: www.python.org
- SQLAlchemy: www.sqlalchemy.org
- Codecademy Python course (free): www.codecademy.com/tracks/python

Questions?

EPA United States Environmental Protection Agency