Data Analysis Tools for Quality Assurance

Assessing the Utility of the R-Programming Language in Technical Systems Audits

Doug Jager

US EPA Region 4 Science and Ecosystem Support Division

Abstract

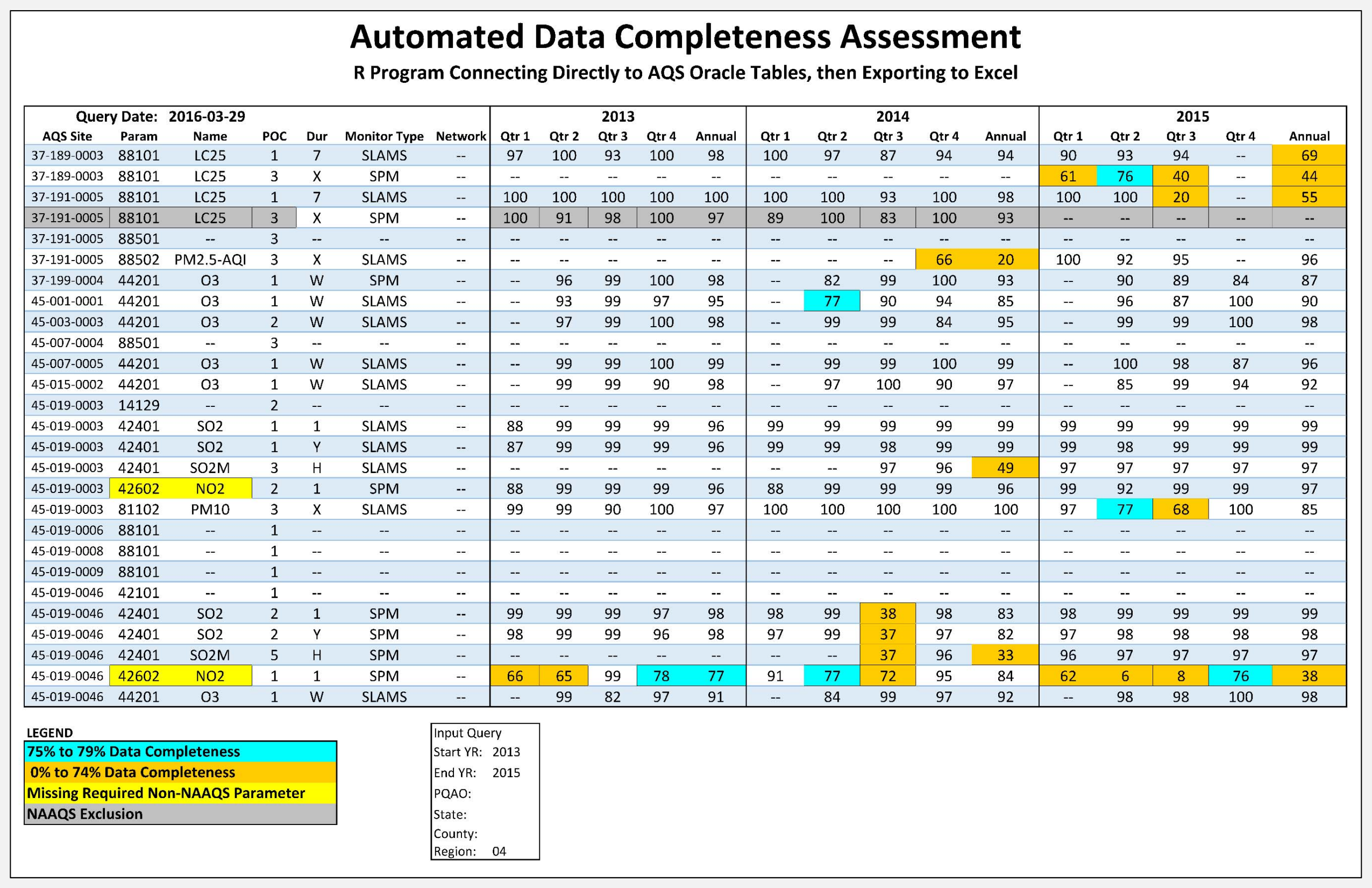
With decreasing resources and increasing demands being placed on local, state, tribal, and federal governments, improving the efficiency and effectiveness of data quality assessments is becoming ever more critical to conducting a successful Technical Systems Audit (TSA). Automated data analysis tools serve to drive consistency in data evaluations, enhance the speed in performing data reviews, and for TSAs, liberate limited staffing resources for other high value activities in the audit.

Introduction

EPA Region 4 SESD has evaluated the R programming language for both its effectiveness in performing data reduction and automating data quality assessments as well as assessing the learning curve required for developing analysis tools for data quality assessments. This poster provides a quick overview demonstrating how the R programming language can be used to assist in TSAs. Examples presented here show how the R programming language can be used to visualize data in quality assessments and how data summaries can be easily exported to other office products such as MS Excel™.

Directly Accessing AQS Oracle Tables for Automating Data Quality Assessments

Two R programs have been developed as demonstration projects for accessing data in AQS. These assessment tools examine Data Completeness and Network Summaries at Regional, State, County, and Primary Quality Assurance Organization (PQAO) levels. Both R programs directly access the Oracle Tables in AQS via the RODBC R Package, summarize and analyze the results, and then export the analyzed datasets into formatted Excel™ spreadsheets using the XLConnect R package. Both programs were designed such that the user requires no knowledge of Structured Query Language (SQL) and only requires a basic familiarity with the R interpreter and interface package such as RStudio™.



Automated Data Completeness Assessment

R Program Connecting Directly to AQS Oracle Tables, then Exporting to Excel

Query Date: 2016-03-29							2013					2014					2015				
AQS Site	Param	Name	POC	Dur	Monitor Type	Network	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Annual
37-189-0003	88101	LC25	1	7	SLAMS	--	97	100	93	100	98	100	97	87	94	94	90	93	94	--	69
37-189-0003	88101	LC25	3	X	SPM	--	--	--	--	--	--	--	--	--	--	--	61	76	40	--	44
37-191-0005	88101	LC25	1	7	SLAMS	--	100	100	100	100	100	100	100	93	100	98	100	100	20	--	55
37-191-0005	88101	LC25	3	X	SPM	--	100	91	98	100	97	89	100	83	100	93	--	--	--	--	--
37-191-0005	88501	--	3	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
37-191-0005	88502	PM2.5-AQI	3	X	SLAMS	--	--	--	--	--	--	--	--	--	66	20	100	92	95	--	96
37-199-0004	44201	O3	1	W	SPM	--	--	96	99	100	98	--	82	99	100	93	--	90	89	84	87
45-001-0001	44201	O3	1	W	SLAMS	--	--	93	99	97	95	--	77	90	94	85	--	96	87	100	90
45-003-0003	44201	O3	2	W	SLAMS	--	--	97	99	100	98	--	99	99	84	95	--	99	99	100	98
45-007-0004	88501	--	3	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
45-007-0005	44201	O3	1	W	SLAMS	--	--	99	99	100	99	--	99	99	100	99	--	100	98	87	96
45-015-0002	44201	O3	1	W	SLAMS	--	--	99	99	90	98	--	97	100	90	97	--	85	99	94	92
45-019-0003	14129	--	2	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
45-019-0003	42401	SO2	1	1	SLAMS	--	88	99	99	99	96	99	99	99	99	99	99	99	99	99	99
45-019-0003	42401	SO2	1	Y	SLAMS	--	87	99	99	99	96	99	99	98	99	99	99	98	99	99	99
45-019-0003	42401	SO2M	3	H	SLAMS	--	--	--	--	--	--	--	--	97	96	49	97	97	97	97	97
45-019-0003	42602	NO2	2	1	SPM	--	88	99	99	99	96	88	99	99	99	96	99	92	99	99	97
45-019-0003	81102	PM10	3	X	SLAMS	--	99	99	90	100	97	100	100	100	100	100	97	77	68	100	85
45-019-0006	88101	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
45-019-0008	88101	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
45-019-0009	88101	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
45-019-0046	42101	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
45-019-0046	42401	SO2	2	1	SPM	--	99	99	99	97	98	98	99	38	98	83	98	99	99	99	99
45-019-0046	42401	SO2	2	Y	SPM	--	98	99	99	96	98	97	99	37	97	82	97	98	98	98	98
45-019-0046	42401	SO2M	5	H	SPM	--	--	--	--	--	--	--	--	37	96	33	96	97	97	97	97
45-019-0046	42602	NO2	1	1	SPM	--	66	65	99	78	77	91	77	72	95	84	62	6	8	76	38
45-019-0046	44201	O3	1	W	SLAMS	--	--	99	82	97	91	--	84	99	97	92	--	98	98	100	98

LEGEND

- 🟦 75% to 79% Data Completeness
- 🟧 0% to 74% Data Completeness
- 🟨 Missing Required Non-NAAQS Parameter
- ⬜ NAAQS Exclusion

Input Query	
Start YR:	2013
End YR:	2015
PQAO:	
State:	
County:	
Region:	04

Table Above (*Automated Data Completeness Assessment*):

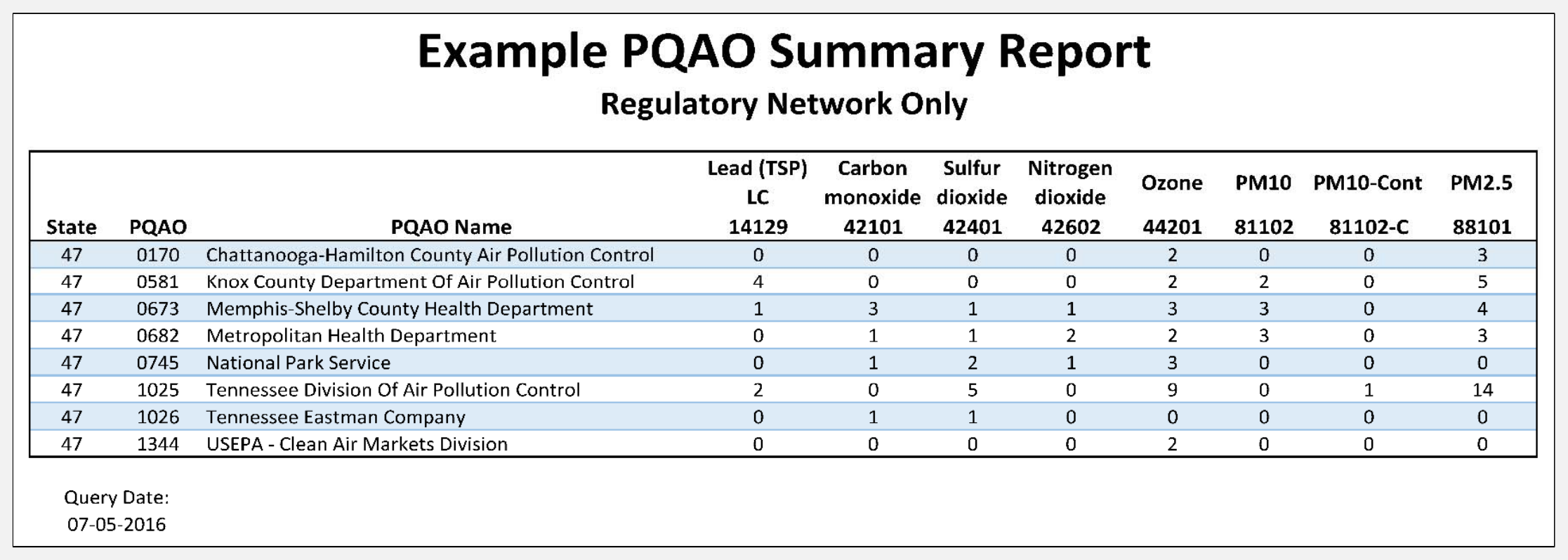
The Assessment Tool extracts quarterly and annual data completeness statistics from AQS. Multiple years of data can be queried at once. The data is exported from R to Excel™ with most formatting automatically performed by the assessment tool; this includes cell highlighting, merging cells, cell boarders, custom title and subtitle, adding of legend, and column header labeling. The exported spreadsheet includes tabs for Data Completeness that is Site Sorted and Parameter Sorted, tabs listing Inactive Monitors and Inactive Monitors that have never reported data, as well as a tab for the raw unformatted data so these results can be easily imported by other databases and data analysis software.

In addition to identifying quarters and years where low data capture occurred, the assessment tool detects NAAQS Excluded monitors, and highlights criteria analyzers that are not reporting required non-criteria parameters (i.e., NO & NO_X channels from NO_2 analyzers and 5-minute SO_2 measurements for SO_2 analyzers).

Tables to the Right (*Example PQAO Summary Report* and *Example Site Summary Report*):

These tables are similar to monitoring network summaries that are often found in Annual Network Plans. These crosstab tables greatly improve the process of verifying that the network summary tables found in the Annual Network Plans match the site and monitoring records stored in AQS. These reports can also be used to verify that PQAOs with non-regulatory monitors have their monitors appropriately NAAQS Excluded.

***Example Site Summary Report*:** In addition to the features common to both reports, this crosstab table allows the auditors to quickly determine the make and models of analyzers comprising a monitoring network. This information can be used to ensure that SOPs are established for all makes and models of analyzers being employed. The R program automatically generates the legend that provides a description of the analyzer for the AQS Method Codes used in the table.

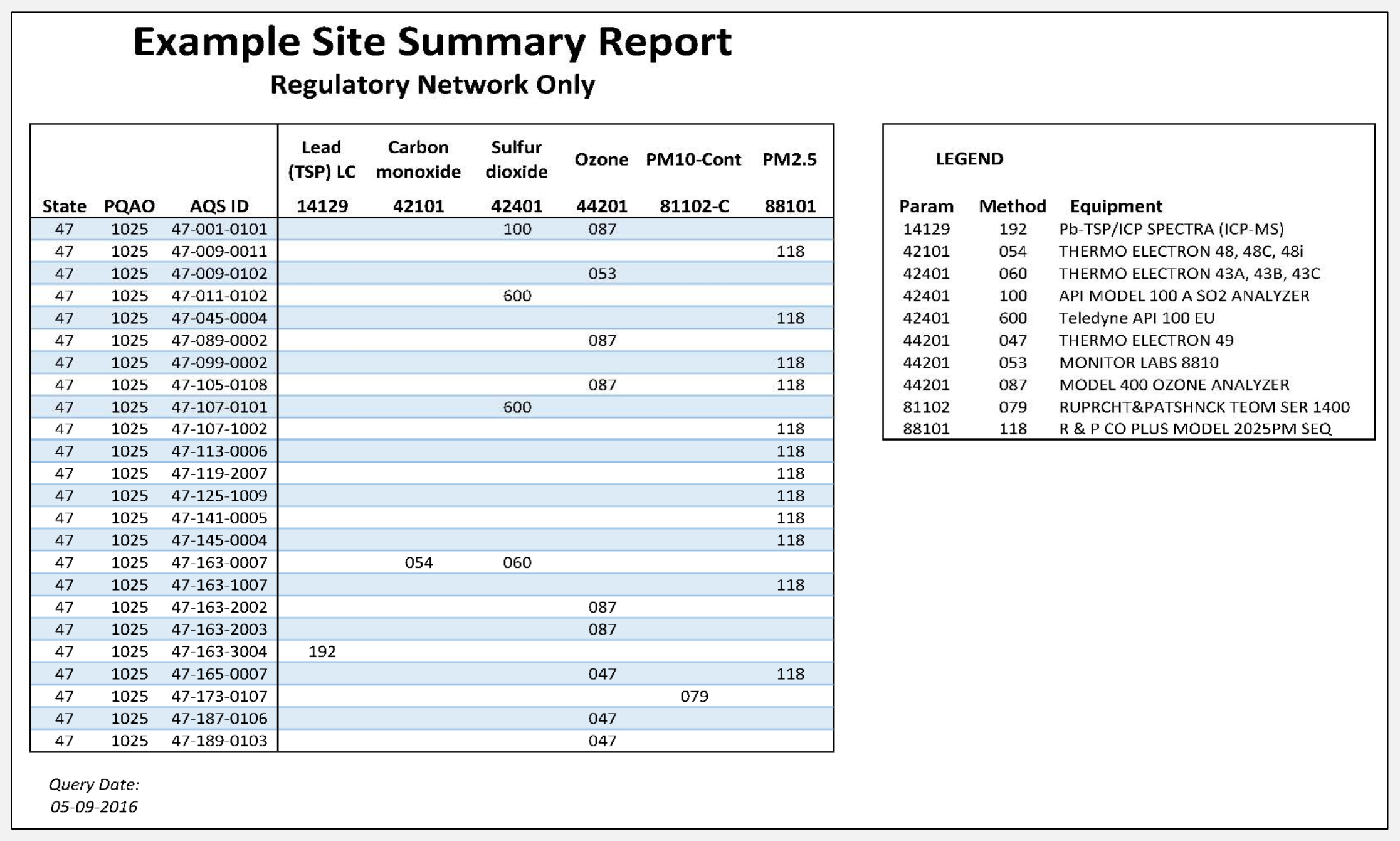


Example PQAO Summary Report

Regulatory Network Only

			Lead (TSP) LC	Carbon monoxide	Sulfur dioxide	Nitrogen dioxide	Ozone	PM10	PM10-Cont	PM2.5
State	PQAO	PQAO Name	14129	42101	42401	42602	44201	81102	81102-C	88101
47	0170	Chattanooga-Hamilton County Air Pollution Control	0	0	0	0	2	0	0	3
47	0581	Knox County Department Of Air Pollution Control	4	0	0	0	2	2	0	5
47	0673	Memphis-Shelby County Health Department	1	3	1	1	3	3	0	4
47	0682	Metropolitan Health Department	0	1	1	2	2	3	0	3
47	0745	National Park Service	0	1	2	1	3	0	0	0
47	1025	Tennessee Division Of Air Pollution Control	2	0	5	0	9	0	1	14
47	1026	Tennessee Eastman Company	0	1	1	0	0	0	0	0
47	1344	USEPA - Clean Air Markets Division	0	0	0	0	2	0	0	0

Query Date:
07-05-2016



Example Site Summary Report

Regulatory Network Only

			Lead (TSP) LC	Carbon monoxide	Sulfur dioxide	Ozone	PM10-Cont	PM2.5
State	PQAO	AQS ID	14129	42101	42401	44201	81102-C	88101
47	1025	47-001-0101			100	087		
47	1025	47-009-0011						118
47	1025	47-009-0102				053		
47	1025	47-011-0102			600			
47	1025	47-045-0004						118
47	1025	47-089-0002				087		
47	1025	47-099-0002						118
47	1025	47-105-0108				087		118
47	1025	47-107-0101			600			
47	1025	47-107-1002						118
47	1025	47-113-0006						118
47	1025	47-119-2007						118
47	1025	47-125-1009						118
47	1025	47-141-0005						118
47	1025	47-145-0004						118
47	1025	47-163-0007		054	060			
47	1025	47-163-1007						118
47	1025	47-163-2002				087		
47	1025	47-163-2003				087		
47	1025	47-163-3004	192					
47	1025	47-165-0007				047		118
47	1025	47-173-0107					079	
47	1025	47-187-0106				047		
47	1025	47-189-0103				047		

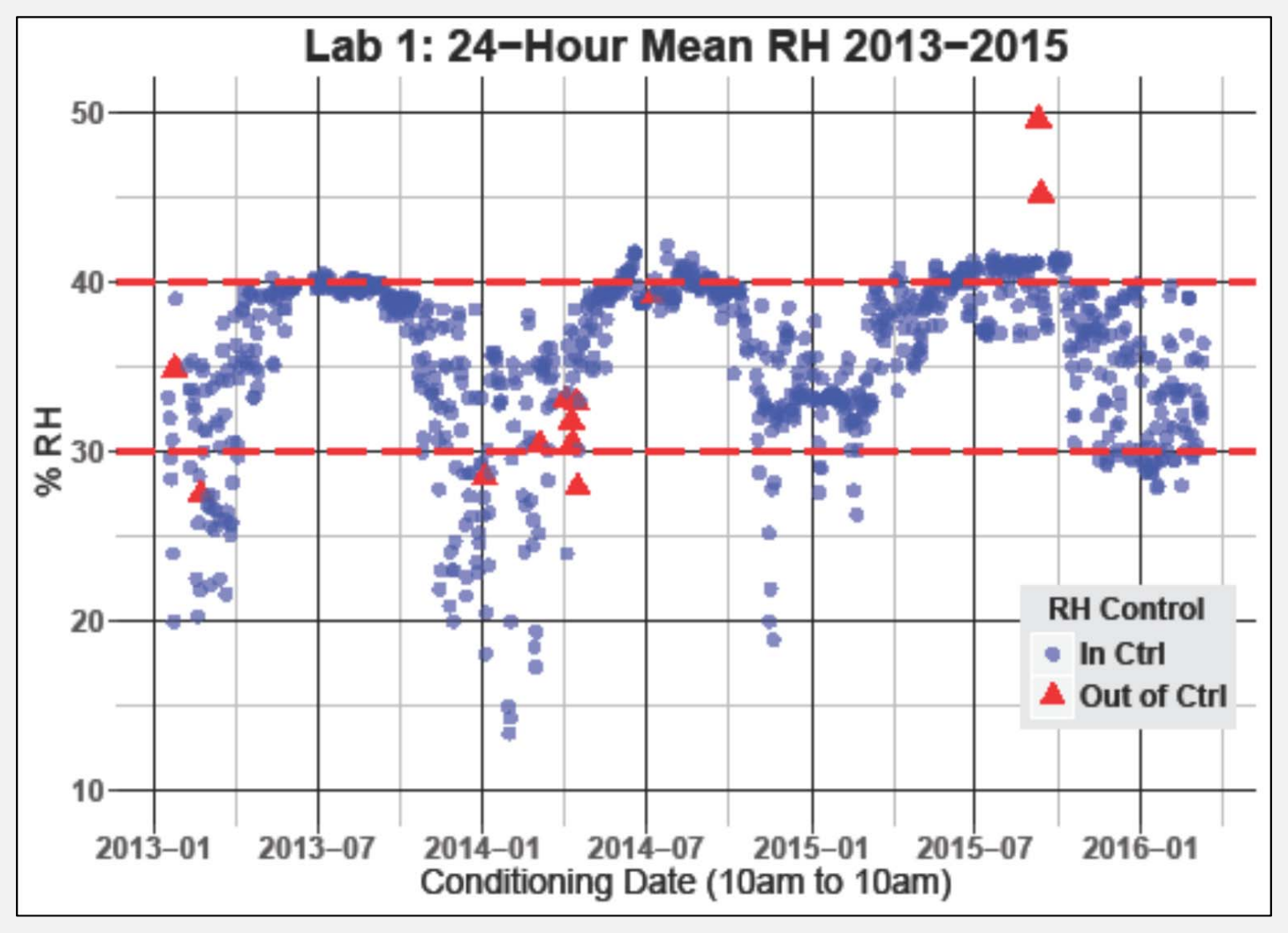
LEGEND

Param	Method	Equipment
14129	192	Pb-TSP/ICP SPECTRA (ICP-MS)
42101	054	THERMO ELECTRON 48, 48C, 48i
42401	060	THERMO ELECTRON 43A, 43B, 43C
42401	100	API MODEL 100 A SO2 ANALYZER
42401	600	Teledyne API 100 EU
44201	047	THERMO ELECTRON 49
44201	053	MONITOR LABS 8810
44201	087	MODEL 400 OZONE ANALYZER
81102	079	RUPRCHT&PATSHNCK TEOM SER 1400
88101	118	R & P CO PLUS MODEL 2025PM SEQ

Query Date:
05-09-2016

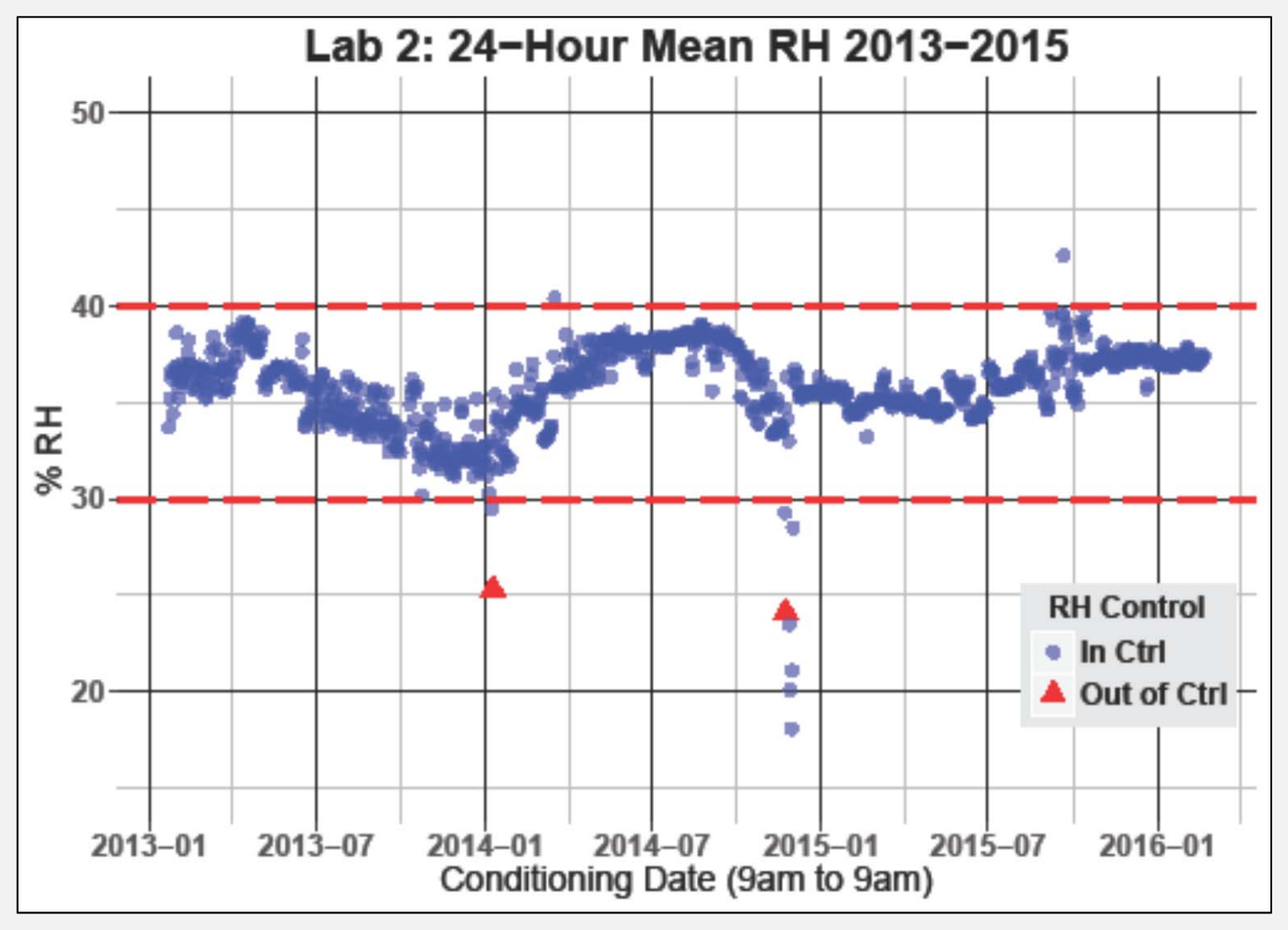
Control Charting: Filter Conditioning Performance in Gravimetric Labs.

Several air monitoring programs operate $PM_{2.5}$ gravimetric laboratories in EPA Region 4. TSAs in recent years have found QA/QC concerns at some of these laboratories. At the time of these TSAs, the Region 4 auditors did not have these data visualization tools to assist in diagnosing the performance of the Lab.'s filter conditioning processes. TSA auditors had to rely on manually spot checking records which is time intensive and does not provide a comprehensive conceptual QA model of the laboratory's performance. To address this deficiency, EPA staff developed a visualization tool using the R programming language. The below figures illustrate the effectiveness of R for visualizing and analyzing very large datasets efficiently and quickly. File formats for the Labs were CSV, Tab Delimited, and MS Access™. In some cases, the minute records were stored in 100's of files. The R program used for these control charts was found to be easily adaptable to evaluate data generated from multiple proprietary laboratory formats.



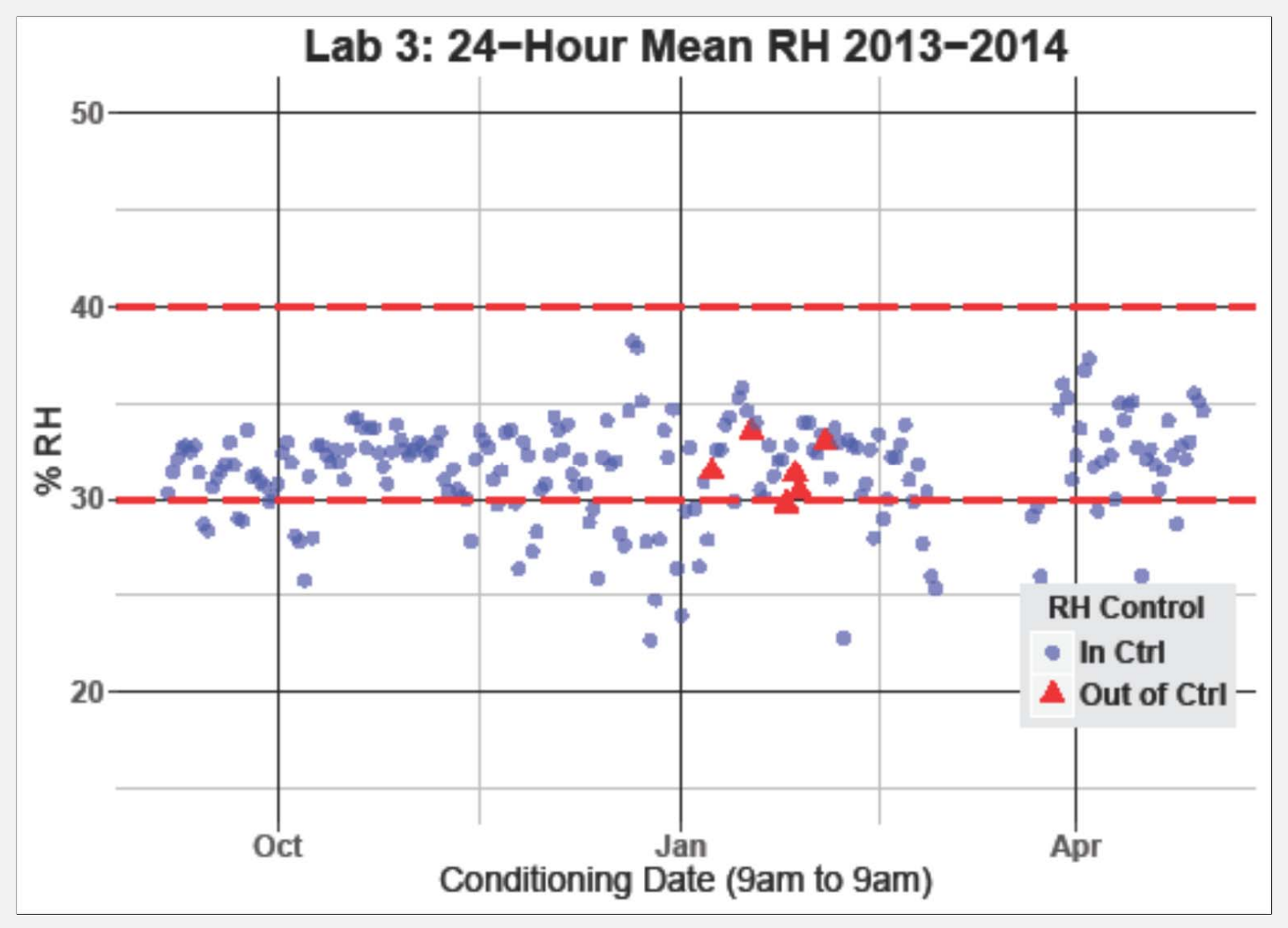
Lab 1: HVAC was unable to maintain humidity in winter and summer. Some RH control issues. Potential Pre/Post conditioning issues due to seasonal "ramping and cycling" of RH.

Daily Means summarized from minute data stored as:
150 Tab Delimited files 238,425 Records



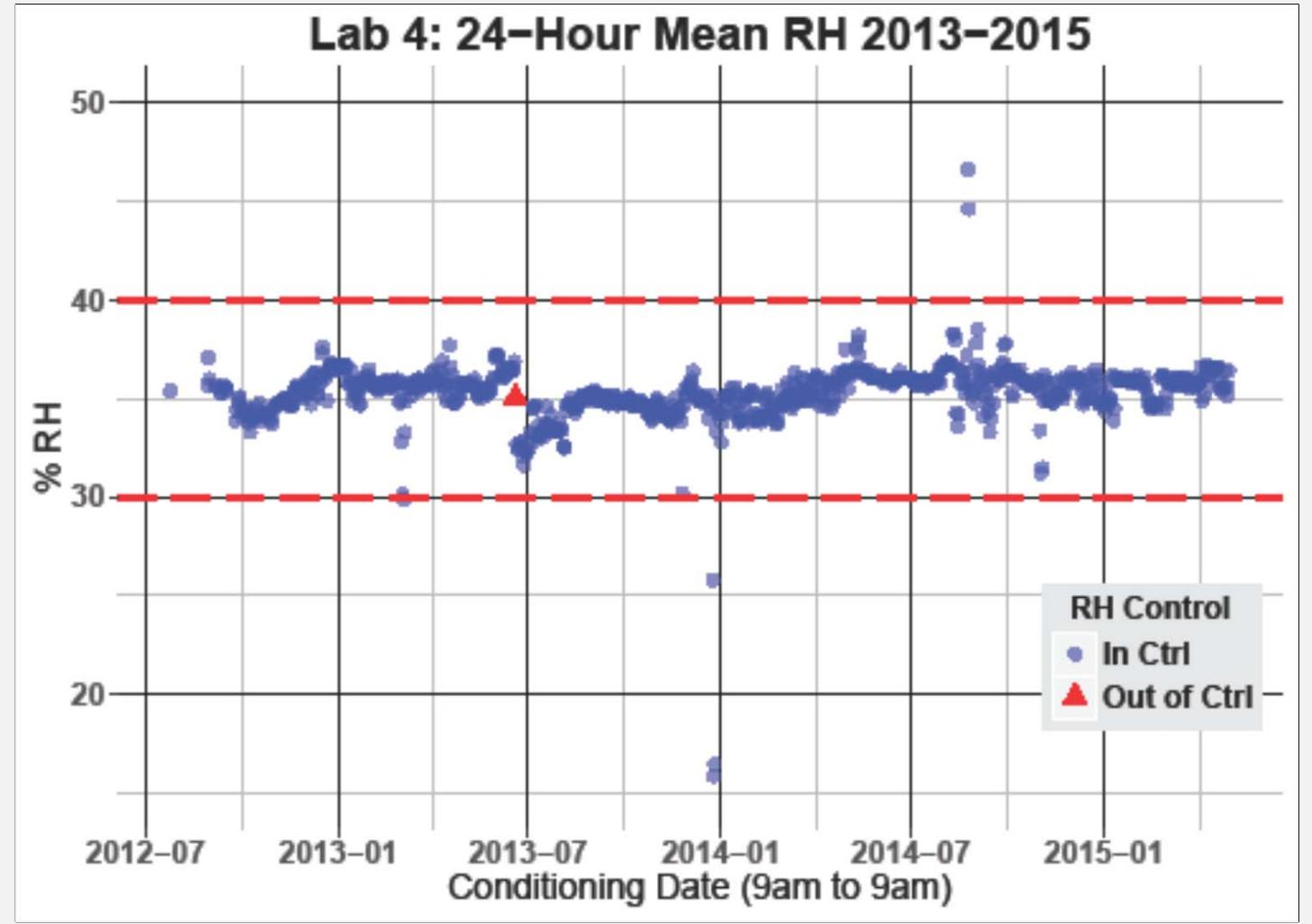
Lab 2: Grav. Lab correctly conditions $PM_{2.5}$ filters. Excursions beyond method requirements were infrequent, quickly detected by quality control, and remedied immediately.

Daily Means summarized from minute data stored as:
528 Tab Delimited files 7,467,869 Records



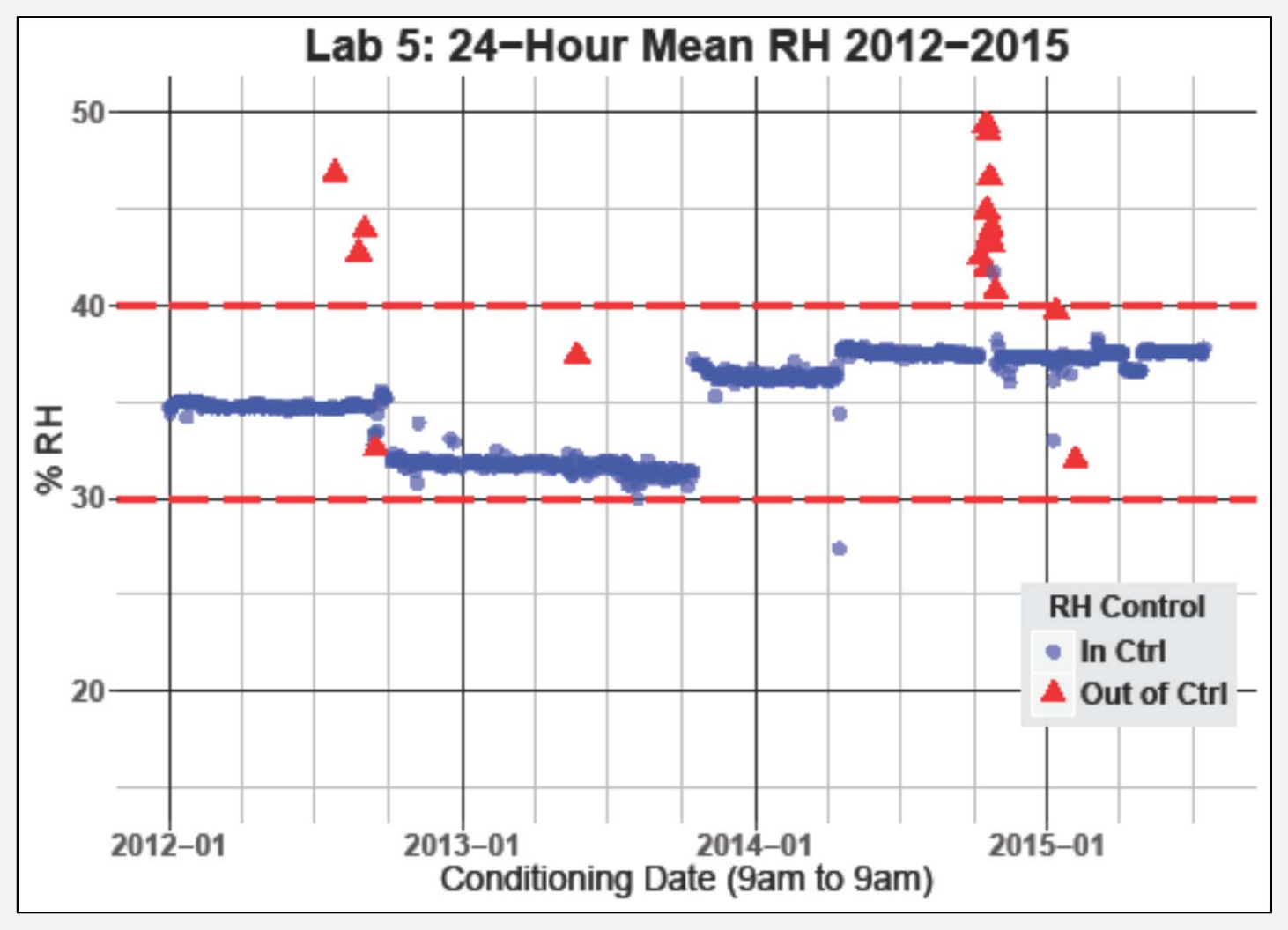
Lab 3: HVAC was unable to maintain humidity in target range (systematically low RH). Some RH control issues. Potential Pre/Post RH conditioning issues due to a lack of precision for the mean RH.

Daily Means summarized from minute data stored as:
131 Tab Delimited files 1,187,978 Records



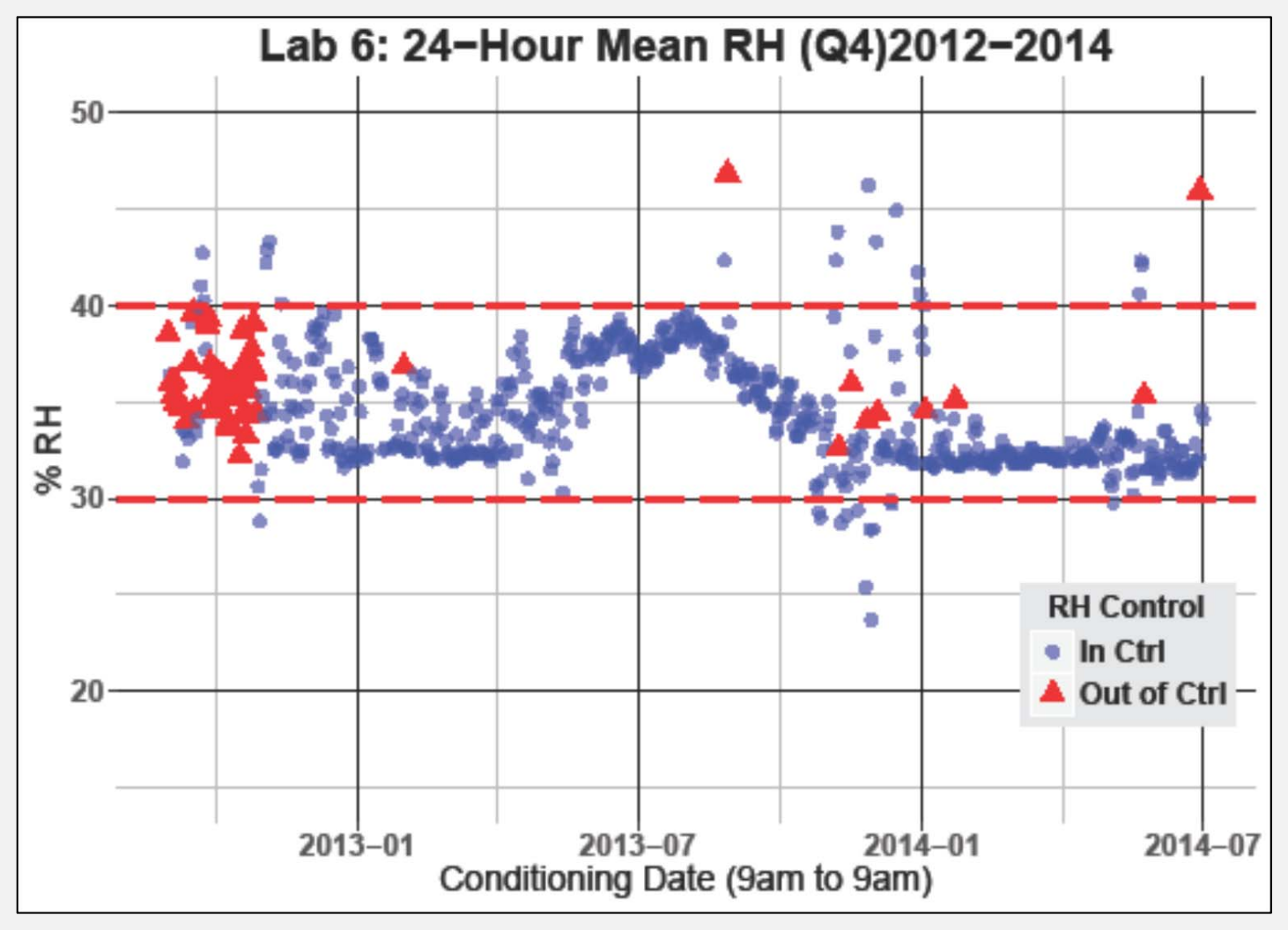
Lab 4: Environmental Controls were diagnostically healthy. HVAC was able to maintain humidity in both winter and summer. Lab was only rarely out of spec. and was quickly brought back into control when QC was not met.

Daily Means summarized from minute data stored as:
1 MS Access DB file 1,093,951 Records



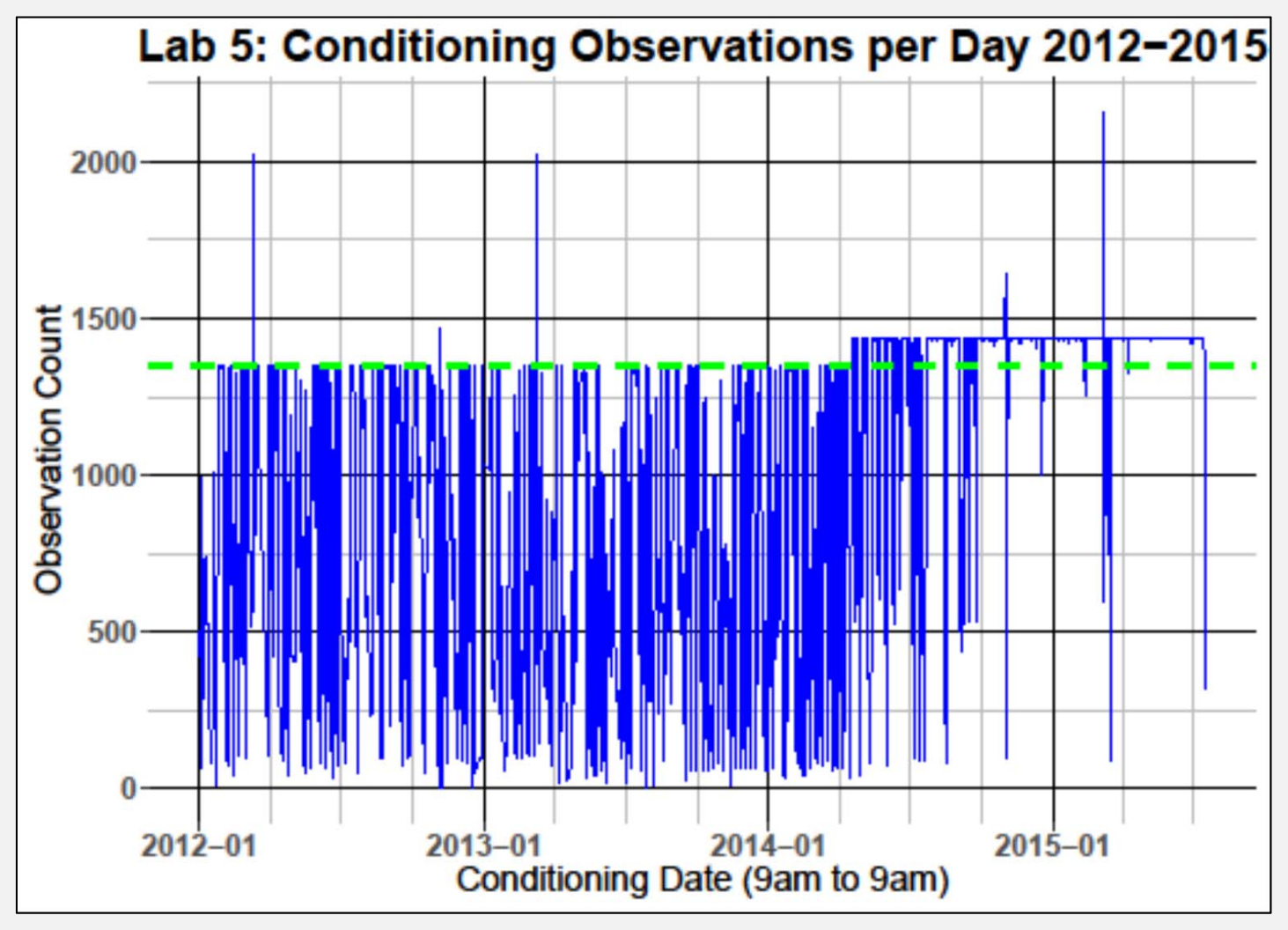
Lab 5: Some RH Control Issues. Control Chart appears to show Lab. was diagnostically healthy. However, RH means were often not valid 24-hr means, but were often based on only a few hours of minute readings. See Figure directly below for examination of observations in each mean.

Daily Means summarized from minute data stored as:
1 CSV file 1,090,849 Records

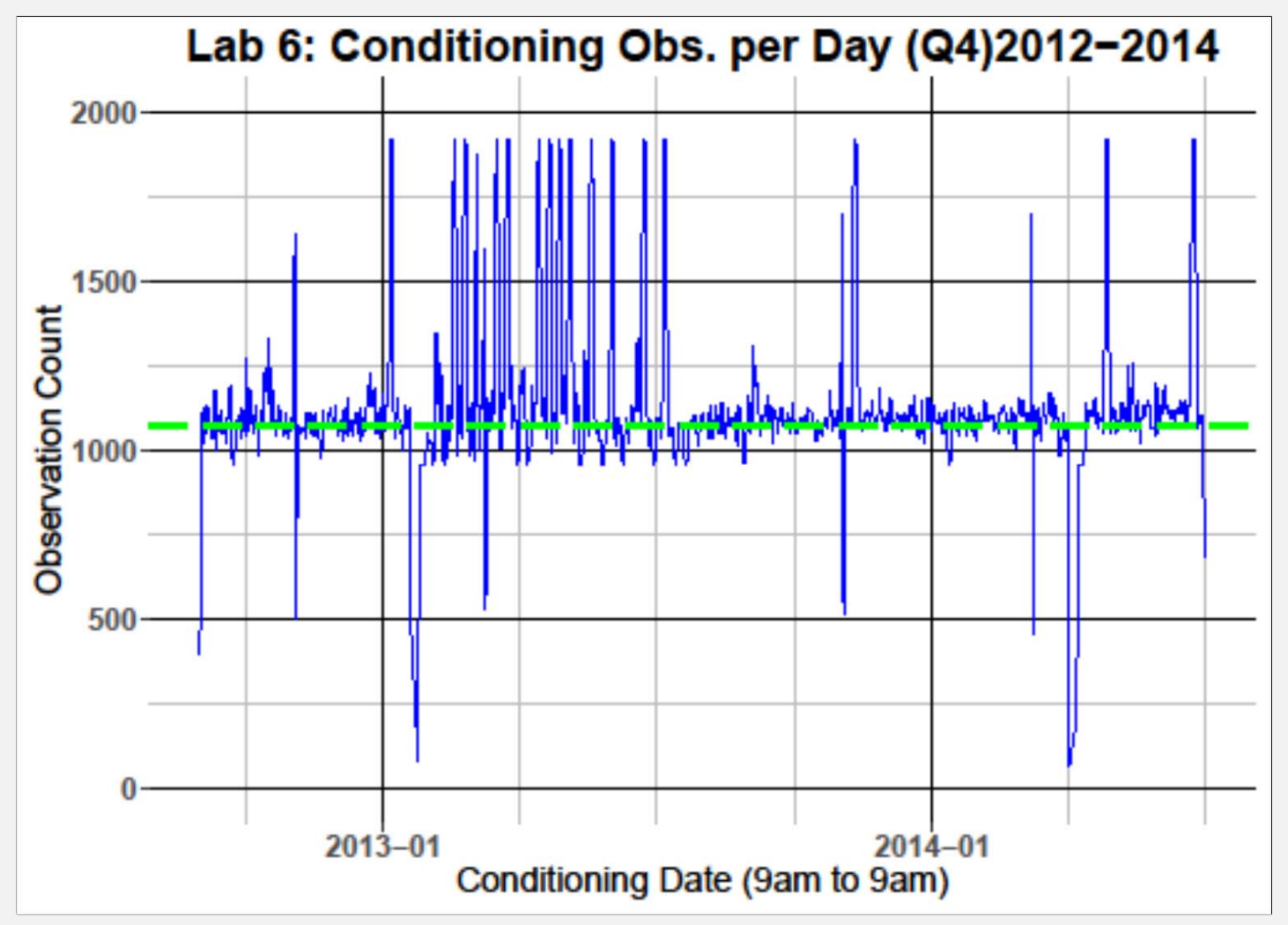


Lab 6: RH Control issues are evident for first 3 months in control chart, large blocks of data were impacted. Occurrences of daily 24-hr Mean RH both above and below the regulatory limits for filter conditioning were present.

Daily Means summarized from minute data stored as:
8 CSV files 750,709 Records



Lab 5: Loss of Single (LOS) was a routine occurrence for the data logger recording the environmental conditions. Mean RH and Temp. were not based on a complete 24-hr period until late 2014.



Lab 6: Good and complete logging of data. Observing "double counting" of minute readings is not uncommon. Only infrequent Loss of Single (LOS).

Conclusions

The automated data analysis tools developed for this demonstration project were found to enhance the speed in performing data quality assessments and have the potential to improve consistency for performing audits of data quality. The learning curve associated with the R programming language is not insignificant, but programming in R has not been as difficult as originally anticipated.

In addition to the examples provided here, Data Analysis tools are in development for:

- Reconciling single point precision & audit results with routine ambient air measurements,
- Evaluating Collocation requirements by PQAO,
- Reconciling flowrate verifications & audit results with routine PM ambient air measurements,
- Evaluating routine measurements in AQS that are potentially impacted by test atmospheres,
- Ensuring routine measurement results are bracketed by QA/QC Checks.

Looking forward, it is hoped that these data quality assessment tools can be exported to state, tribal, and local ambient air programs for use in their quality assurance assessments and annual data certifications.

Nick Mangus: The Data Completeness and PQAO Network Summary Reports, and all of the data quality assessment tools currently in development, highly leverage the Oracle Views designed and implemented by Nick Mangus with the US EPA Office of Air Quality Planning & Standards. These Oracle Views to the AQS database have greatly reduced the time required to progress from the conception of an analysis tool to a working prototype ready for evaluation and testing.