Real-time Data Collection and Storage: Best Practices

NEBC Business & The Environment Conference

December 5, 2023

Katheryn Kolesar, PhD

AIR SCIENCES INC.

DENVER • PORTLAND • LOS ANGELES

Overview

STANDARDIZATION FOR DATA COLLECTION AND STORAGE

DIGITAL VS MANUAL QUALITY REVIEW

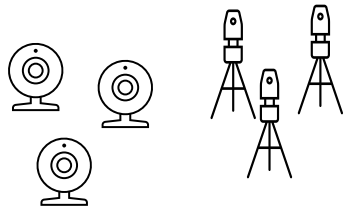
DATA ARCHIVING AND SHARING

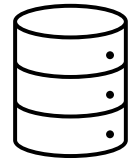
Air Sciences Inc.

- 40 years in air quality permitting and compliance.
- Hard rock mining, lumber mills, transportation, construction, dust mitigation.
- Create custom databases for clients to collect, store, review, archive, and visualize their data.
- Customize third-party databases to import collected data, review, and visualize.
- Interact with publicly available data for quality control and supplementary datasets.
- Augment routine reporting with automated components.

Standardization for Collection and Storage

Database Creation Basics





Acquisition

Method(s) of data logging/telemetry

Variety of sensors/vendors

Field surveys?

Storage

Regulatory drivers

Data types

Data transformations

Geographic information

Dissemination

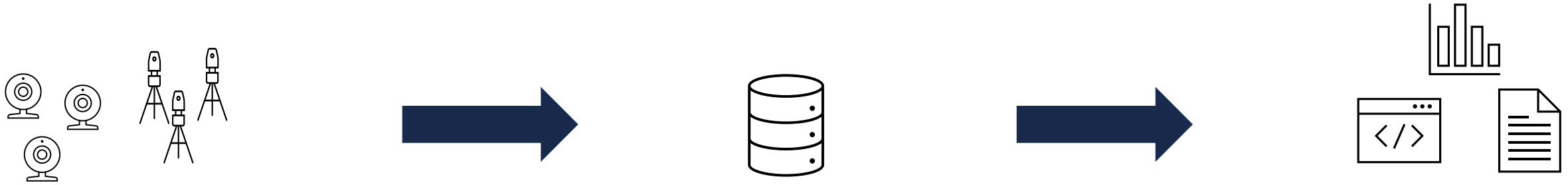
Visualization

Auditing/validation

Alerts/Notifications

Exporting/reporting

Database Creation Basics



Acquisition *Storage* *Dissemination*

Federal Regulatory Monitors		VisionEval	VisionEval
IoT Sensor package (one vendor)		IoT vendor	Custom scripts using API
IoT (many vendors)	APIs to poll data	Eagle.io	Eagle.io Custom scripts
Complex (++monitors, field surveys, images)	Data polling server	Custom DB model + API	Custom web app

Data File Formatting – Common Mistakes

Where was the device?

What time zone?

Pressure? What units?

What are '_E' and '_I'?

Signal, of what?

What is the difference?

What might this be?

	A	B	D	E	K	L	N	O	Q	R	S	T	V	W	X	Y	Z	AA	AB
1	Device_ID	Date_Time	C6H6	CH4	NO2	O2	P	PM10	PM2_5	RH_E	RH_I	Sig	TVOC	TVOC_RAW	T_E	T_I	WBGT	WD	WS
2	240044000150	6/30/2023 0:01	-1	-1	0	-1	842.655	11.074	9.244	-1	66.831	-81.656	0.416	57.159	-1	68.915	-1	107.346	2.426
3	240044000150	6/30/2023 1:01	-1	-1	0	-1	843.813	7.156	6.142	-1	65.521	-81.667	0.405	56.642	-1	69.395	-1	52.735	1.323
4	240044000150	6/30/2023 2:01	-1	-1	0	-1	843.706	5.492	4.767	-1	67.751	-81.65	0.412	56.987	-1	68.638	-1	136.725	2.563
5	240044000150	6/30/2023 3:01	-1	-1	0	-1	844.508	3.781	3.134	-1	67.88	-81.667	0.404	56.681	-1	66.921	-1	216.889	1.426
6	240044000150	6/30/2023 4:01	-1	-1	0	-1	846.27	4.624	3.918	-1	65.855	-81.65	0.396	56.302	-1	66.757	-1	125.872	1.259
7	240044000150	6/30/2023 5:01	-1	-1	0	-1	847.566	6.122	5.151	-1	79.577	-81.683	0.427	57.828	-1	64.526	-1	352.784	3.248
8	240044000150	6/30/2023 6:01	-1	-1	0	-1	848.626	4.332	3.112	-1	96.508	-81.7	0.448	58.949	-1	59.06	-1	358.957	3.726
9	240044000150	6/30/2023 7:01	-1	-1	0	-1	848.601	4.149	3.333	-1	100	-81.8	0.461	59.564	-1	58.925	-1	350.134	2.774
10	240044000150	6/30/2023 8:01	-1	-1	0	-1	848.59	3.072	2.462	-1	98.484	-81.583	0.446	58.817	-1	60.522	-1	310.272	0.846
11	240044000150	6/30/2023 9:01	-1	-1	0	-1	847.972	2.977	2.327	-1	91.362	-81.683	0.434	58.267	-1	60.846	-1	258.793	1.034

data (1)

1 minute after the hour is inconvenient for pairing data in time

-1 is a bad choice for documenting invalid data

Is this valid data?

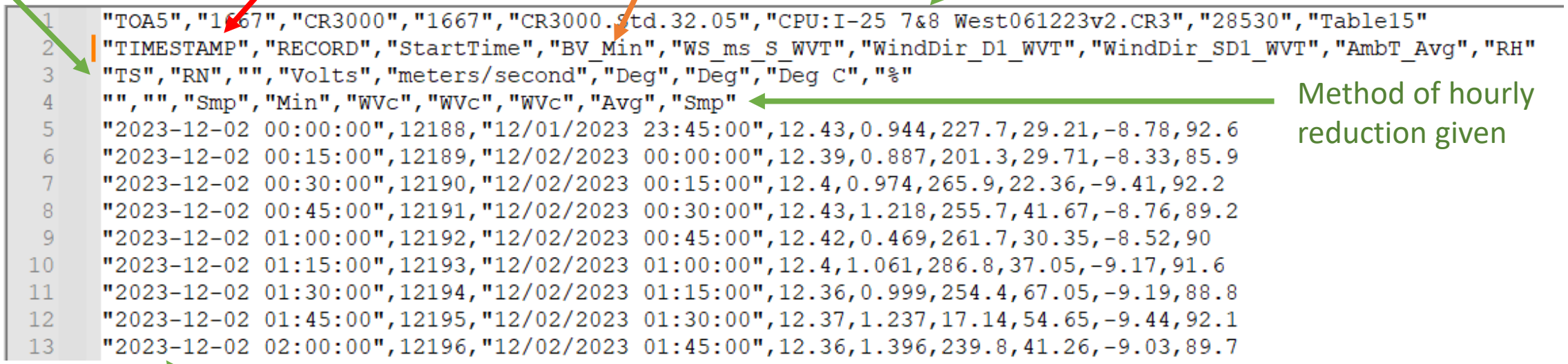
Data File Formatting – Better Example

Units of collection are given

No time zone information

Naming could be more informative

Station name and relative location



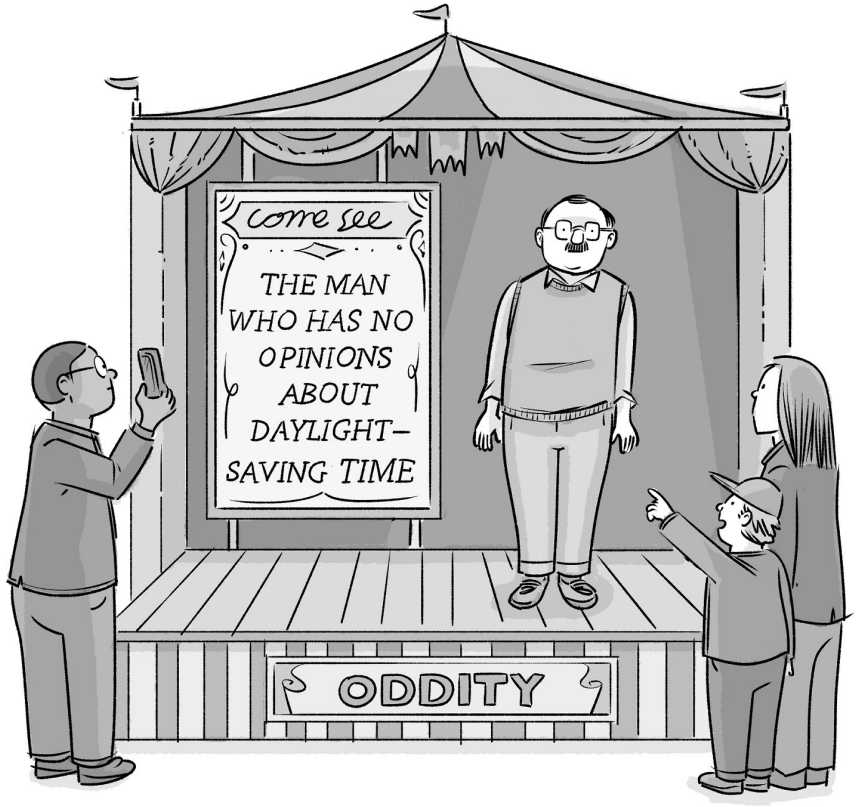
```
1  "TOA5","1667","CR3000","1667","CR3000.Std.32.05","CPU:I-25 7&8 West061223v2.CR3","28530","Table15"
2  "TIMESTAMP","RECORD","StartTime","BV_Min","WS_ms_S_WVT","WindDir_D1_WVT","WindDir_SD1_WVT","AmbT_Avg","RH"
3  "TS","RN","","Volts","meters/second","Deg","Deg","Deg C","%"
4  "","","Smp","Min","WVc","WVc","WVc","Avg","Smp"
5  "2023-12-02 00:00:00",12188,"12/01/2023 23:45:00",12.43,0.944,227.7,29.21,-8.78,92.6
6  "2023-12-02 00:15:00",12189,"12/02/2023 00:00:00",12.39,0.887,201.3,29.71,-8.33,85.9
7  "2023-12-02 00:30:00",12190,"12/02/2023 00:15:00",12.4,0.974,265.9,22.36,-9.41,92.2
8  "2023-12-02 00:45:00",12191,"12/02/2023 00:30:00",12.43,1.218,255.7,41.67,-8.76,89.2
9  "2023-12-02 01:00:00",12192,"12/02/2023 00:45:00",12.42,0.469,261.7,30.35,-8.52,90
10 "2023-12-02 01:15:00",12193,"12/02/2023 01:00:00",12.4,1.061,286.8,37.05,-9.17,91.6
11 "2023-12-02 01:30:00",12194,"12/02/2023 01:15:00",12.36,0.999,254.4,67.05,-9.19,88.8
12 "2023-12-02 01:45:00",12195,"12/02/2023 01:30:00",12.37,1.237,17.14,54.65,-9.44,92.1
13 "2023-12-02 02:00:00",12196,"12/02/2023 01:45:00",12.36,1.396,239.8,41.26,-9.03,89.7
```

Method of hourly reduction given

Two timestamp values, one is interval start time, one is interval end time (default)

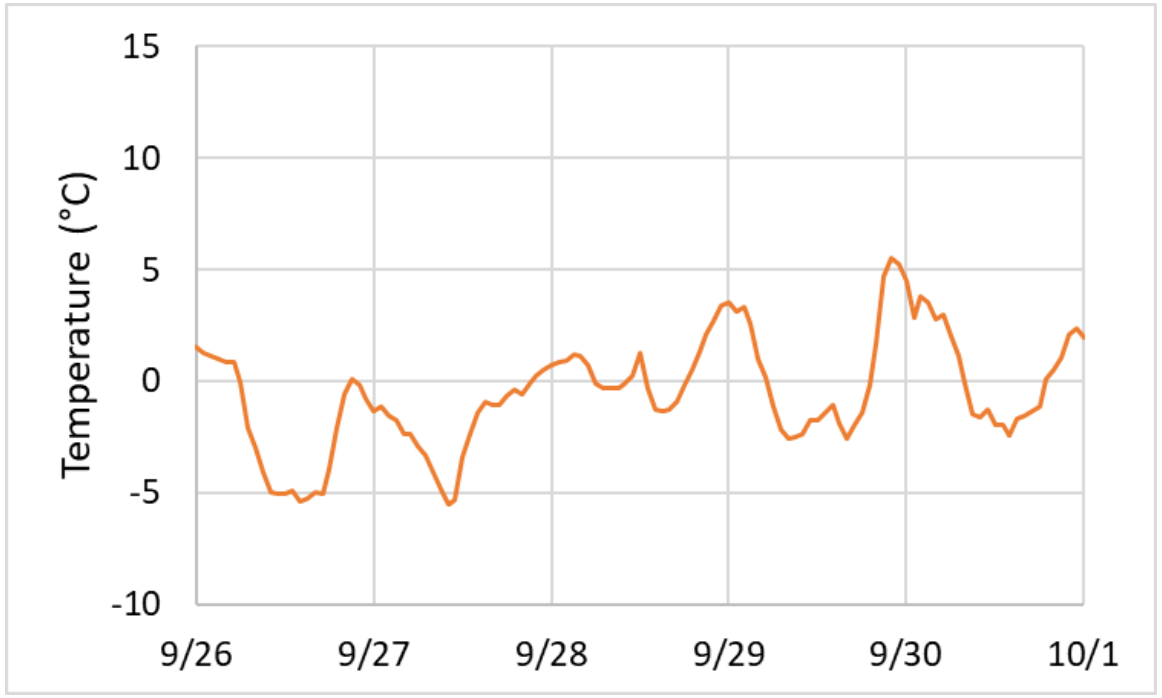
What time is it?

- Time zones
- Time beginning vs Time ending data reduction

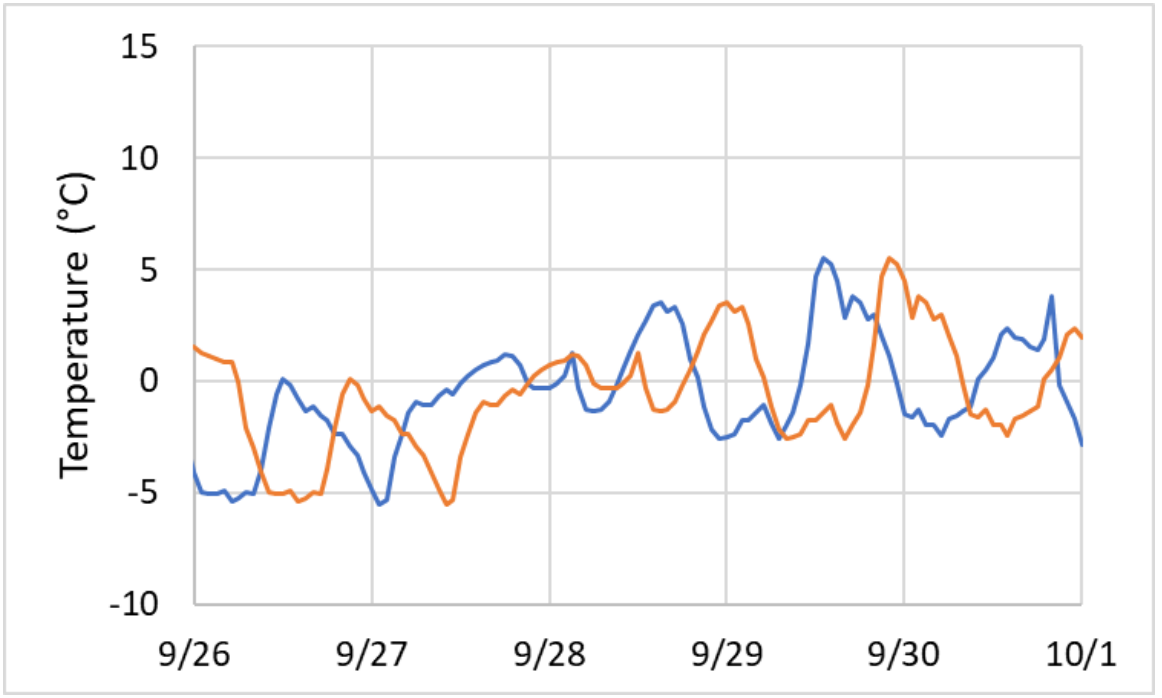


"Technically, Daylight Saving Time isn't time travel, but, sure, I guess if you see another you, try to avoid him."

Time Zones are Important



Datafile incorrectly assumed to use Alaska Standard Time (UTC – 9 hours)



Datafile used Coordinated Universal Time (UTC)

Data Reduction Time Convention

Time Beginning

Timestamp	PM_{10} Concentration (µg/m³)
7/1/2023 1:00	31.833
7/1/2023 1:15	15.153
7/1/2023 1:30	3.71
7/1/2023 1:45	3.721
7/1/2023 2:00	3.781

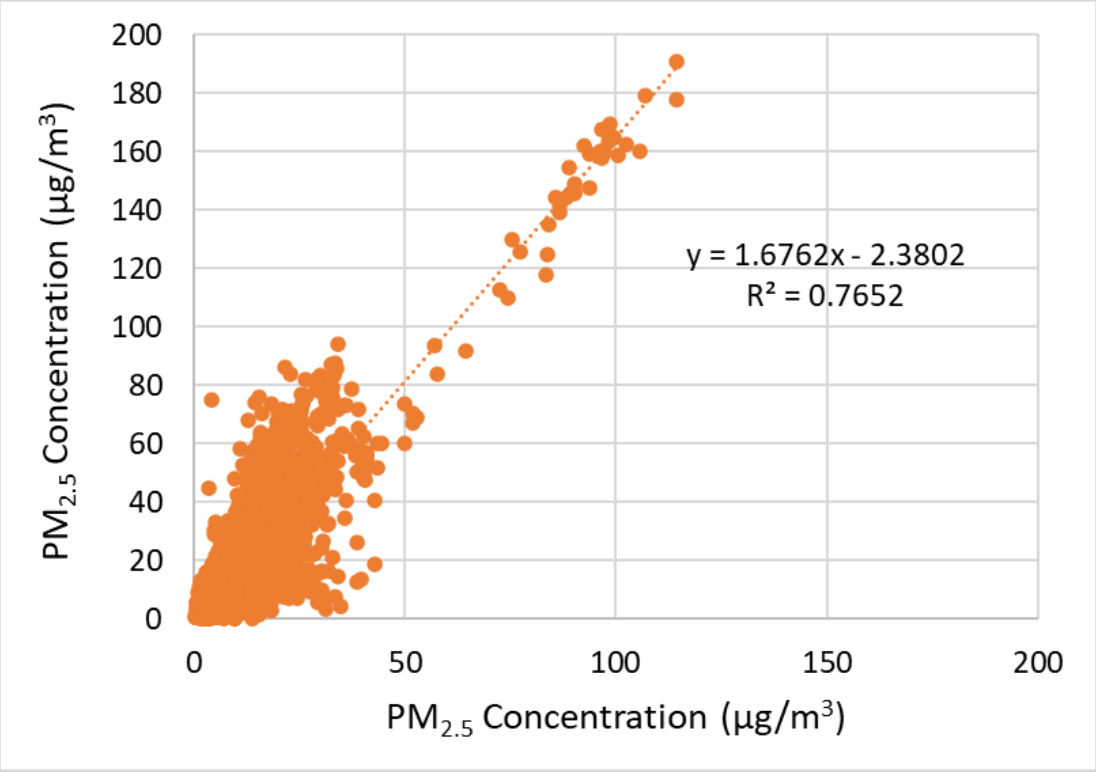
Data gathered from 1:00 to 2:00 are associated with the 1:00 time stamp

Time Ending

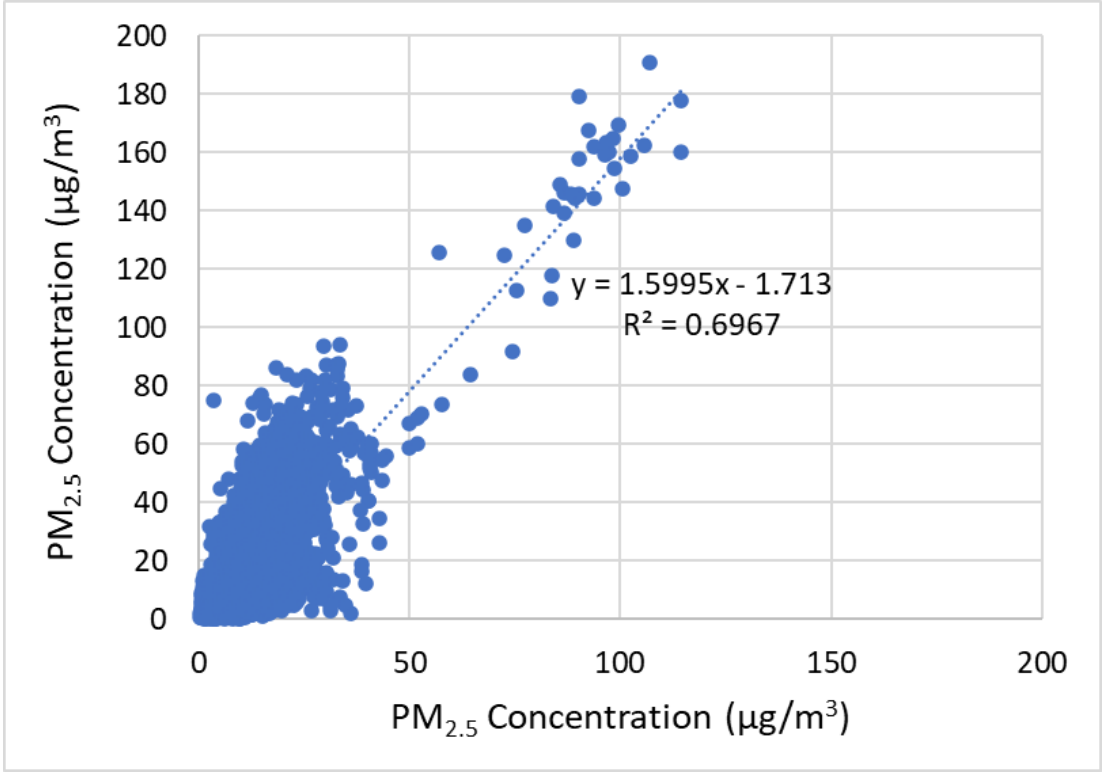
Timestamp	PM_{10} Concentration (µg/m³)
7/1/2023 1:00	31.833
7/1/2023 1:15	15.153
7/1/2023 1:30	3.71
7/1/2023 1:45	3.721
7/1/2023 2:00	3.781

Data gathered from 1:00 to 2:00 are associated with the 2:00 time stamp

What is the big deal?



Data properly matched in time

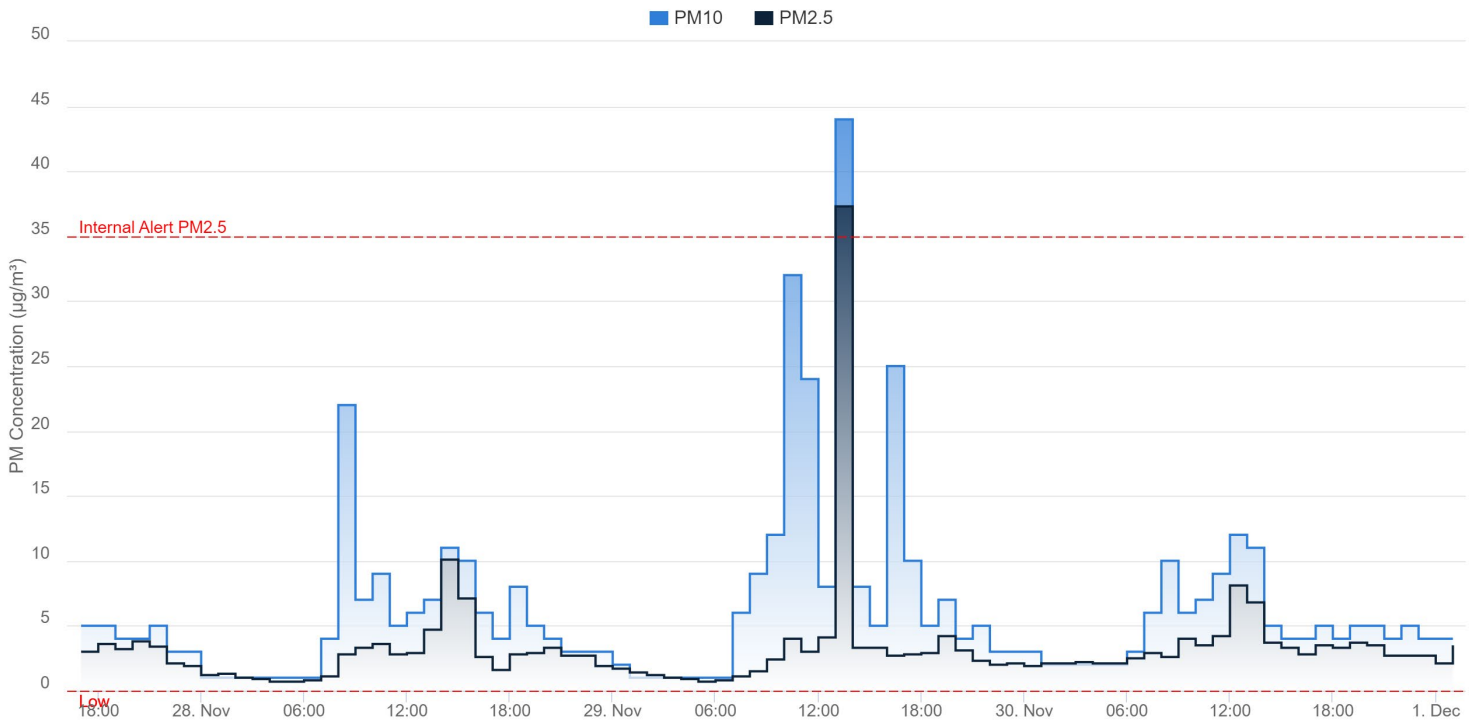


Data on x-axis collected in PDT, Data on y-axis collected in PST

Digital vs Manual Review

Digital Review and Automated Alerts

- Check Instrument Functionality
  - Flow rate out of specification
  - Wind direction does not vary over several hours
- Provide alerts for specific situations
  - Wind speed above a specific threshold
  - Wind direction from a specific sector
  - $PM_{2.5}$ above a threshold



Limits of Digital Review

- Only as good as the user inputs
- Still requires human follow-up and review

Data Archiving & Sharing

File Type Choices to Ensure Accessibility and Shareability of Structured Data

Good Choices

- Delimited Text
- Comma separated values (.csv)
- SQL DDL (.sql)
- Sqlite version 3 (.sqlite)

Okay Choices

- Microsoft Excel (.xlsx)
- OpenOffice (.sxc; .ods)
- DBF (.dbf)
- R (.rdata)
- JSON

Bad Choices

- Microsoft Excel (.xls)
- Microsoft Access (.accdb)
- pdf
- Microsoft Word (.docx)
- MatLab (.mat)
- Other paid program data formats

Summary

AIR SCIENCES INC.

DENVER • PORTLAND • LOS ANGELES

- Database type will depend on project requirements and may evolve over time.
- The ability to invalidate data within the database is important.
- Datafiles need to include metadata.
- Collecting and storing data in UTC time-beginning avoids common pitfalls of real-time data management.
- Good practices for data sharing and archiving will depend on the users and the data.
  - But whatever you do, don't use an .xls file or Microsoft Access