Understanding the relationship between indicators & tracers and vapor intrusion

Dynamic multivariate time series regressions

Riley Mulhern, RTI International

Chris Lutes, Jacobs

Chase Holton, GSI Environmental

AJ Kondash, RTI International

RTI INTERNATIONAL

AEHS Annual Conference | U.S. EPA "State of VI Science" Workshop | March 20-23, 2023

Presentation Overview

- Modeling Aims & Approach
- Data Wrangling & Preprocessing
- Regression Development
- Results
- Conclusions & Next Steps

Modeling Aims & Approach

Challenges with using indicators and tracers to assess VI

- Radon
- Climatic conditions
- Building conditions

?

→

Indoor air VOC concentrations

Long term goals:

- Guide sampling decisions
- Early warnings
- Mitigate exposures
- Soil Gas Safe Communities

Primary aim is to gain insights into the relationship between indicators/ tracer and VI (**not prediction of future VI*)

- Radon
- Climatic conditions
- Building conditions

?

→

Indoor air VOC concentrations

*Need better understanding of the relationship between indicators/tracers and VI **<u>across sites</u>** and **<u>over time</u>** to make **<u>generalizable recommendations</u>**.*

Challenges with using indicators and tracers to assess VI

- Radon
- Climatic conditions
- Building conditions

?

→

Indoor air VOC concentrations

Past work:

- Single variate time series analyses
- Multivariate time series analysis for a single site

Specifying regression predictor variables and outcome variable

Predictor variables

- Radon (R, pCi/L)
- Differential temperature (ΔT, °C)
  - indoor-outdoor
- Differential pressure (ΔP, Pa)
  - positive indicates higher indoor pressure
  - Most sites: indoor-subslab
  - Indianapolis 422 first floor: basement-first floor

Outcome variable

Indoor air VOC concentrations (C, $\mu g/m^3$)

Temporal data sets from three different VI sites across the country

Virginia Site A

- Coastal military site, 120,000 ft^2
- Brick with poured concrete slab 6-8 in. thick
- Separate HVAC zones
- **3 sampling sites** (Office, Supply room, Women's Restroom)
- ~19 months of data
- **Trichloroethylene (TCE)** from historical releases of chlorinated solvents near the site

Sun Devil Manor

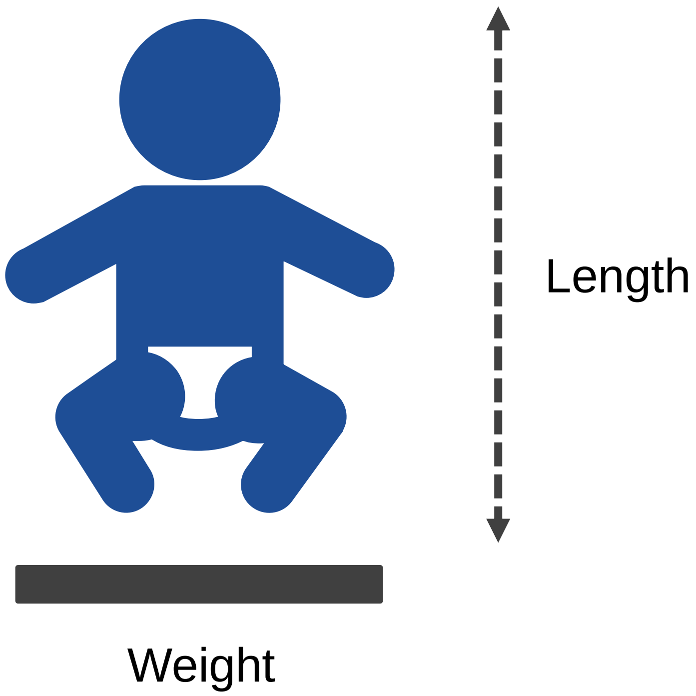
- Layton, UT
- Modern suburban residential
- Split level, ground floor sampling location
- **1 sampling site**
- ~17 months of data
- Trichloroethylene (TCE)

Indianapolis 422

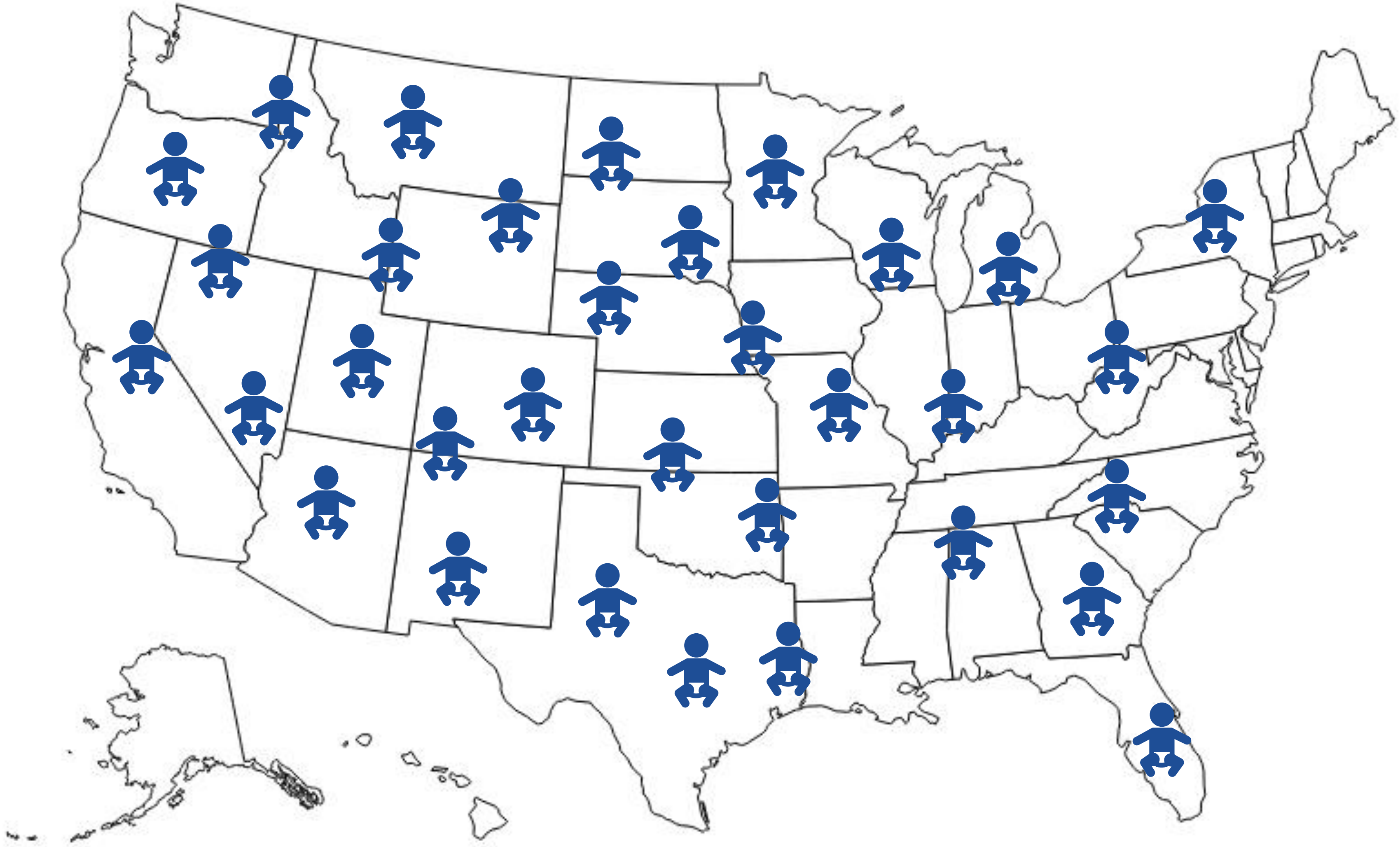
- Indianapolis, IN
- Constructed ~1915
- Duplex
- Wood frame, brick foundation, concrete basement floor
- **2 sampling sites** (Basement, First Floor)
- ~4 months of data
- **Perchloroethylene (PCE)** from historical dry cleaning and adjacent businesses

Ideal data for traditional linear regression: a simple example

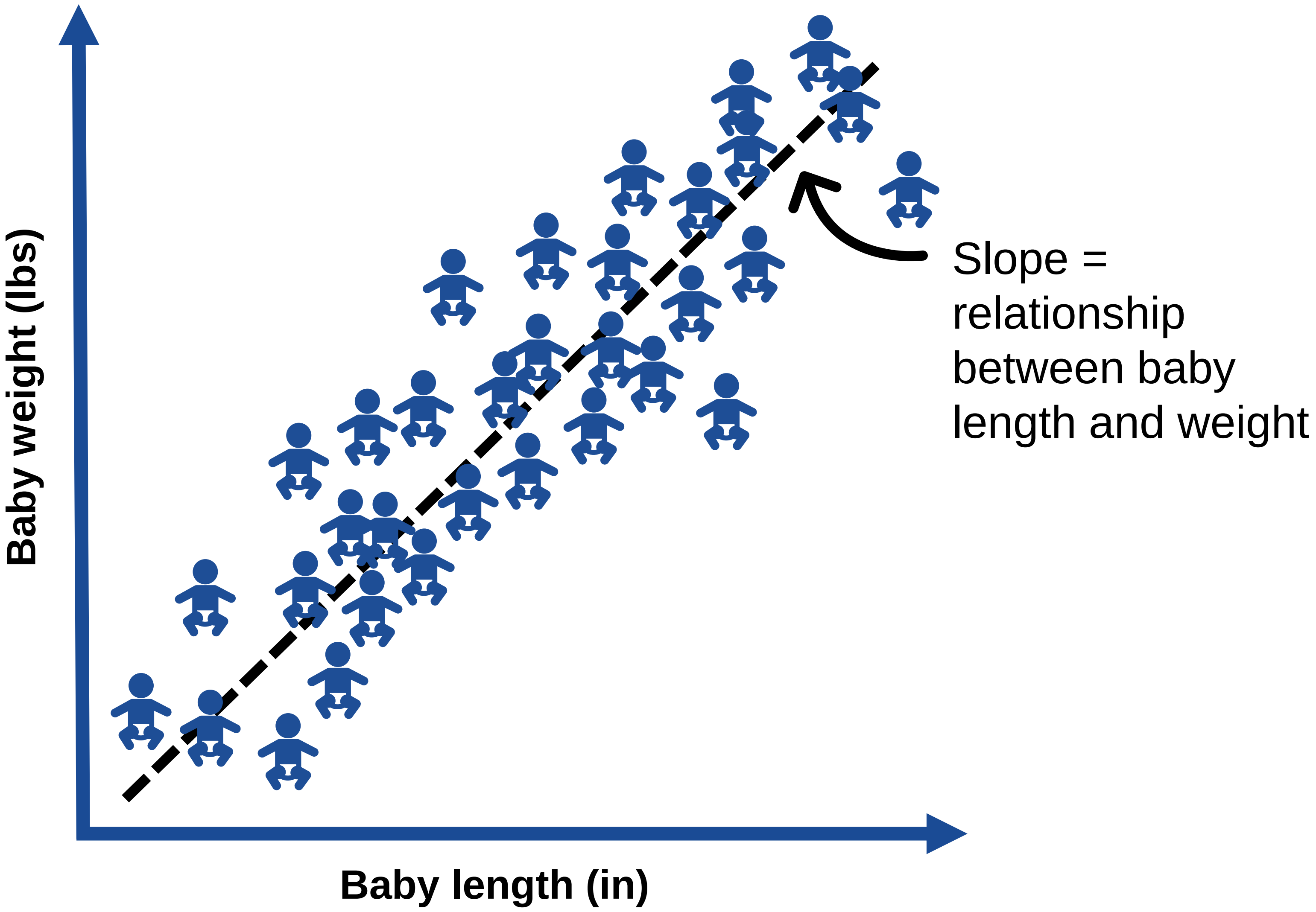
- Assumes all observations are **independent** (not related to each other)
  - Ex. What is the relationship between **length** and **weight** among babies less than 1 year old in the U.S.?
  - Hypothesis: As babies grow longer, their weight increases.
  - Ideal theoretical data set to test this hypothesis using linear regression: single measurements of many individual babies' weights and lengths → **a little bit of data from a lot of different babies**



Ideal data for traditional linear regression: a simple example

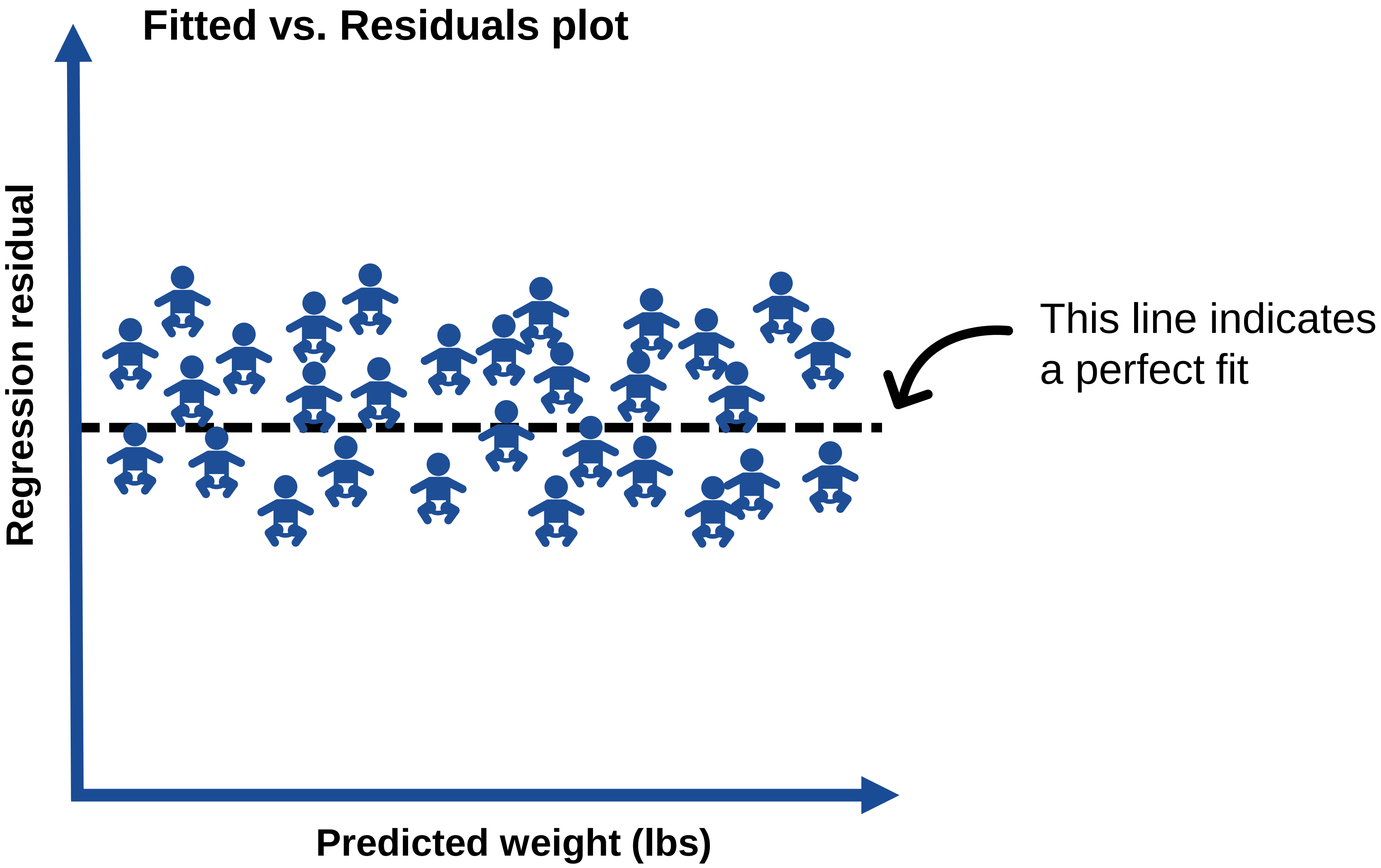


Ideal data for traditional linear regression: a simple example



Ideal data for traditional linear regression: a simple example

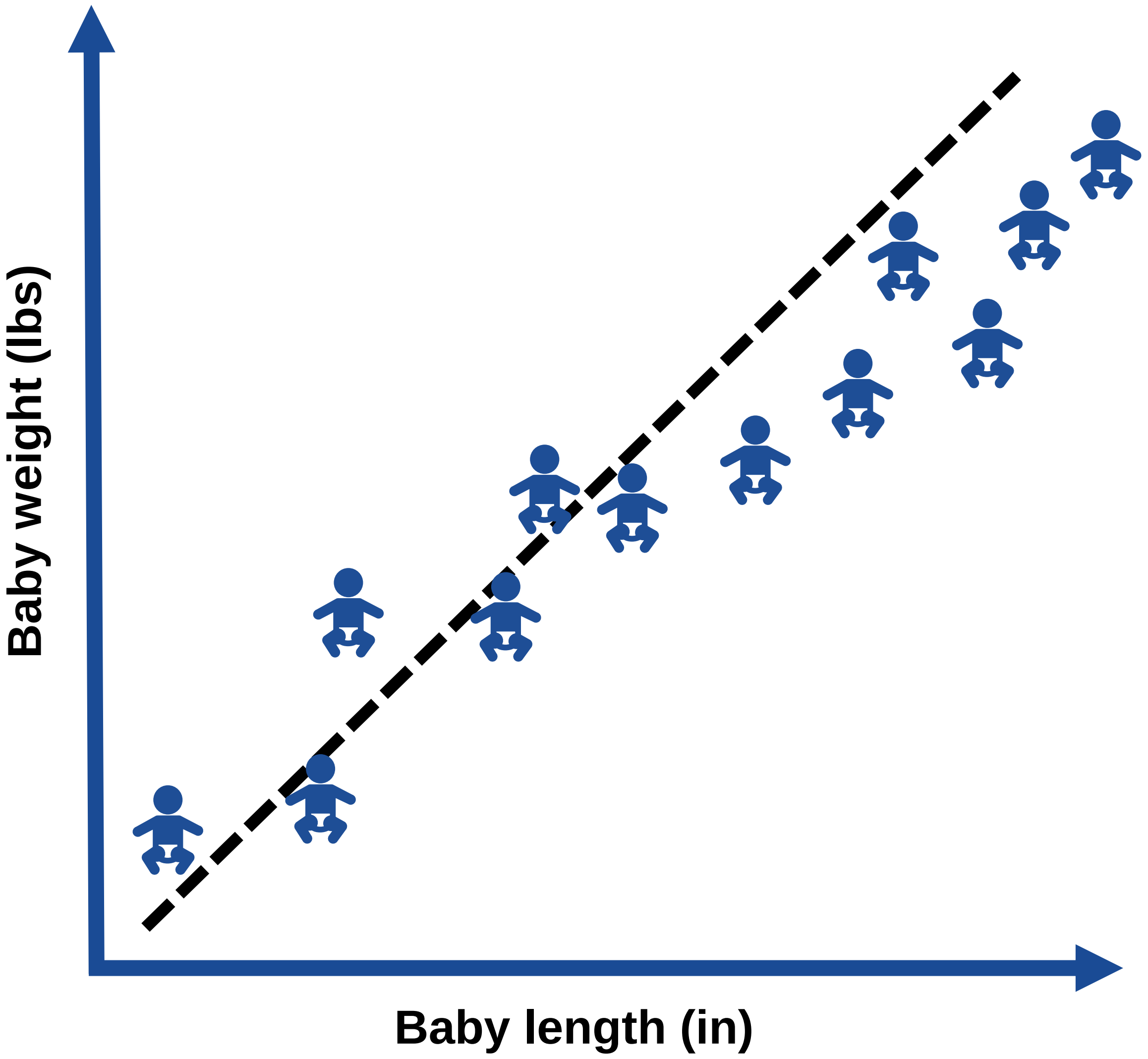
- ✓ **Accurate estimates** of regression coefficients (true relationship between the variables)
- ✓ **High confidence** in the significance of the relationship



Temporal data set: challenges to using traditional regression

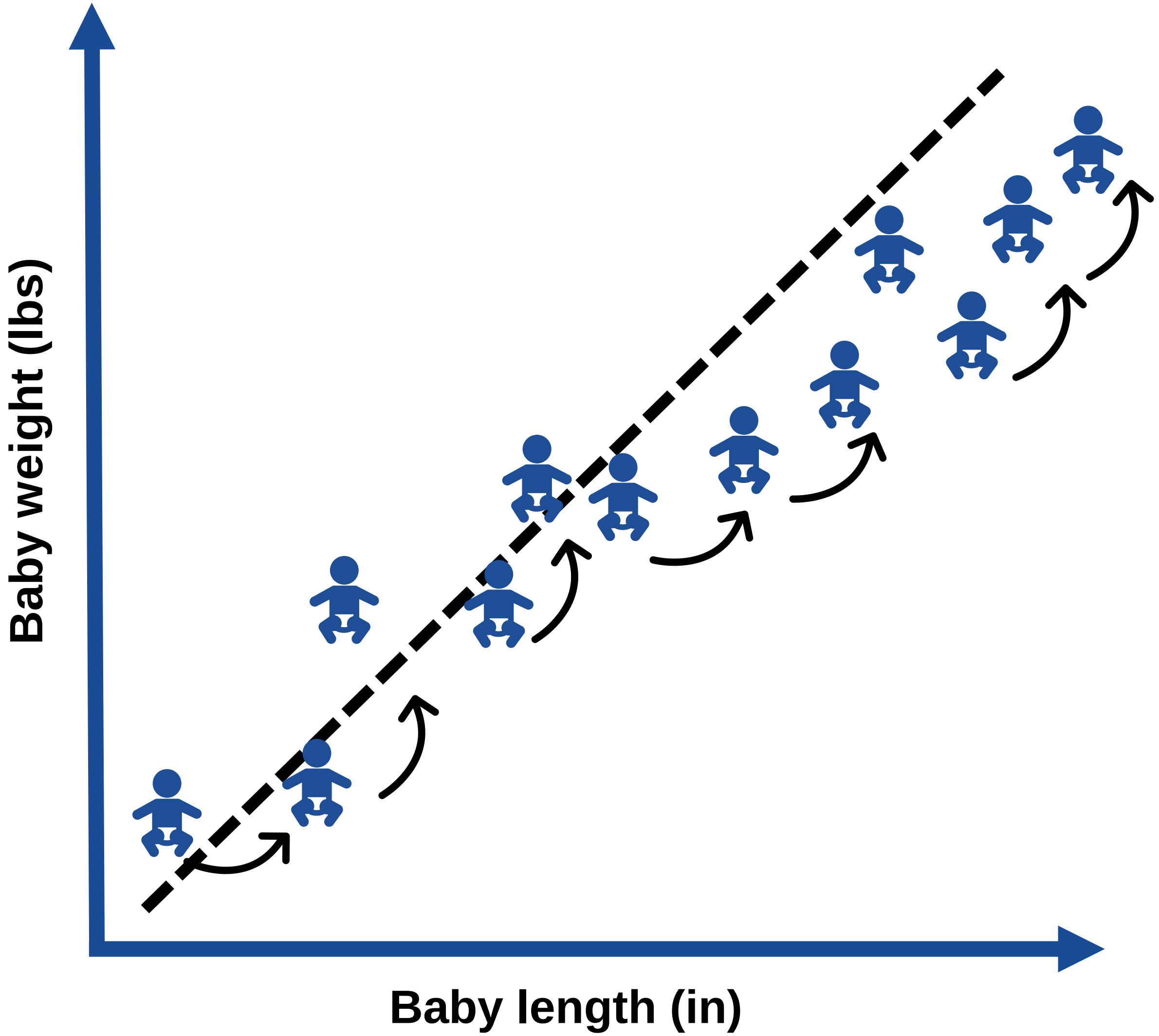
- Actual data set: multiple measurements from a single baby at different points in time → **a lot of data from only one baby**
- Hypothesis: As babies get older, their weight increase.
- Time series data are **not independent** → each measurement is related to measurements in the past

Temporal data set: challenges to using traditional regression



Temporal data set: challenges to using traditional regression

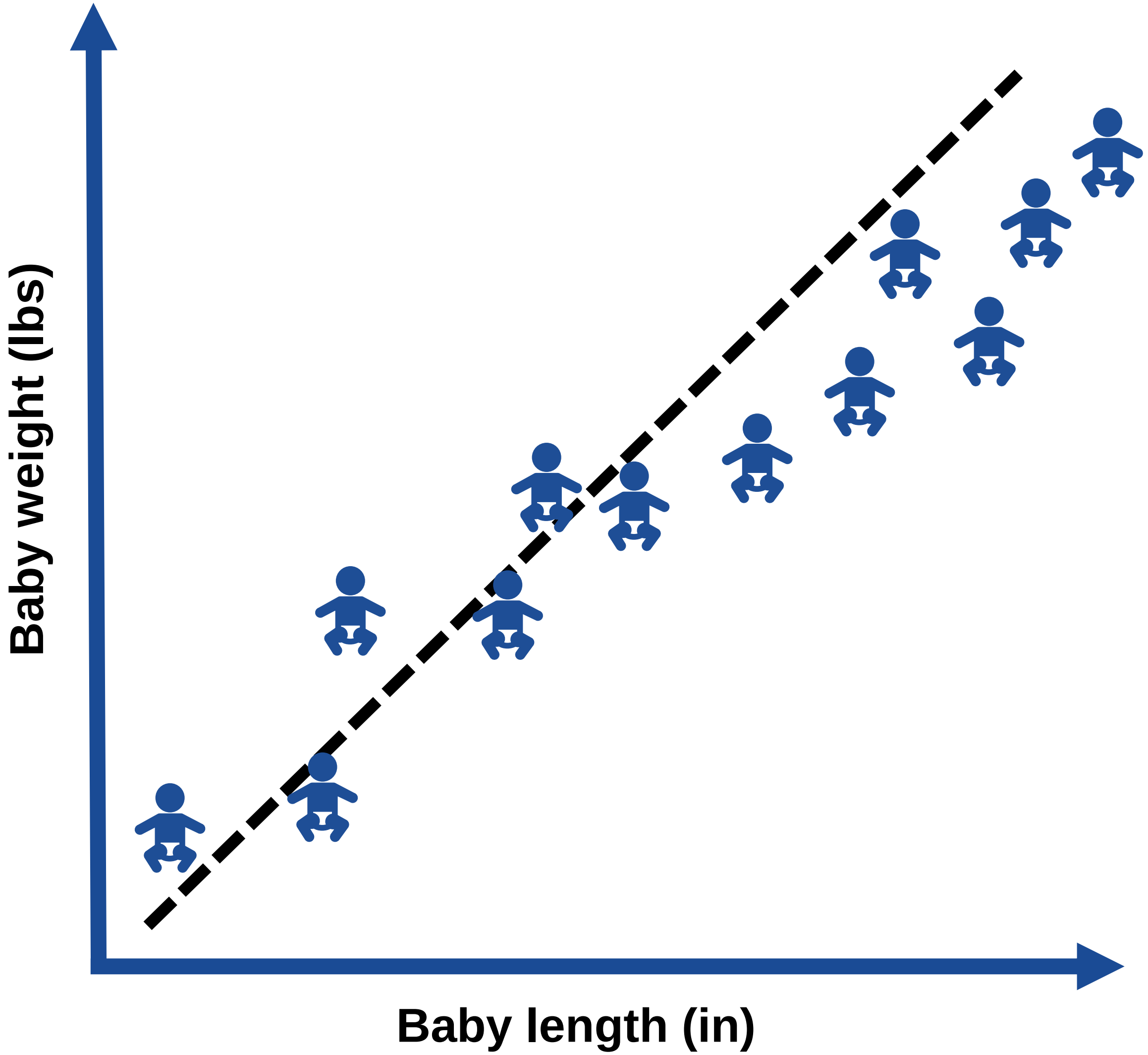
✕ **Autocorrelation**
Leads to inaccurate estimates of the regression coefficients because we ignore important information in the data set.



Temporal data set: challenges to using traditional regression

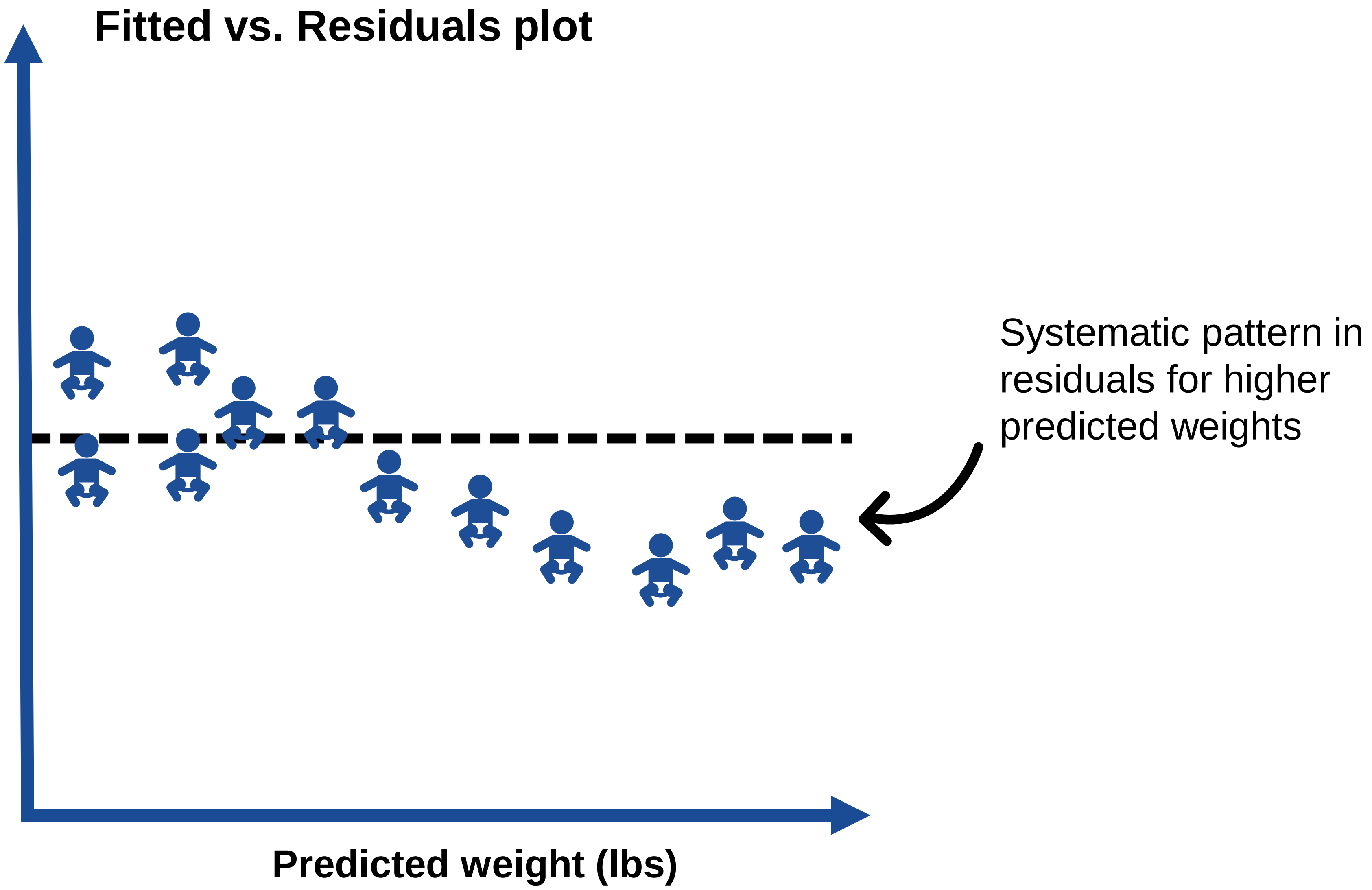
✕ **Spurious regression**
Could conclude that a significant relationship exists between variables when it doesn't actually exist

- Air transport in Australia related to rice production in New Guinea?



Temporal data set: challenges to using traditional regression

✕ **Systematic bias**
Measurement errors over time can lead to systematic bias in predictions

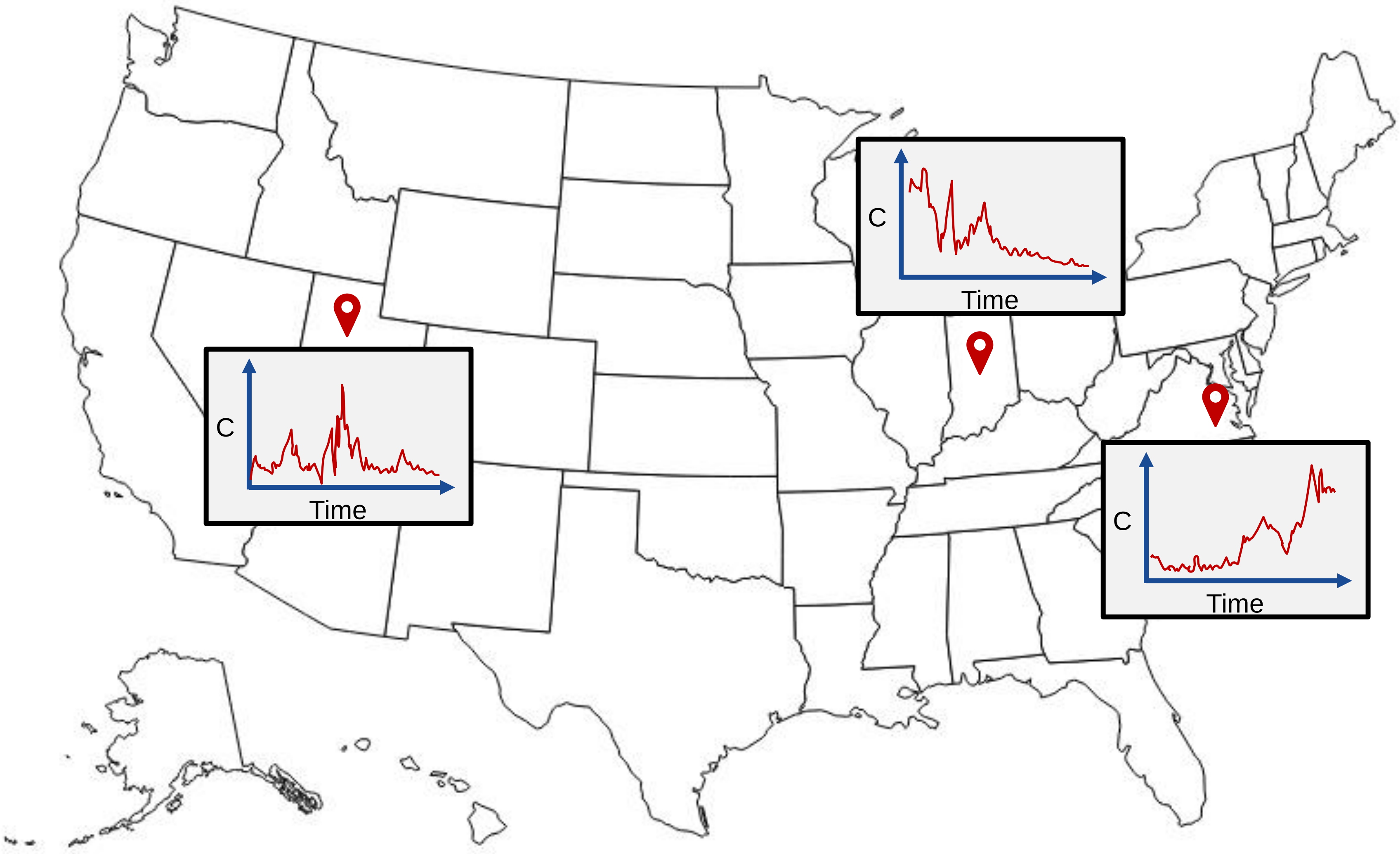


Actual data set: time series data from three sites

- Many observations of VOC concentrations and predictor variables at sequential points in time from three sites across the country

Actual data set: time series data from three sites

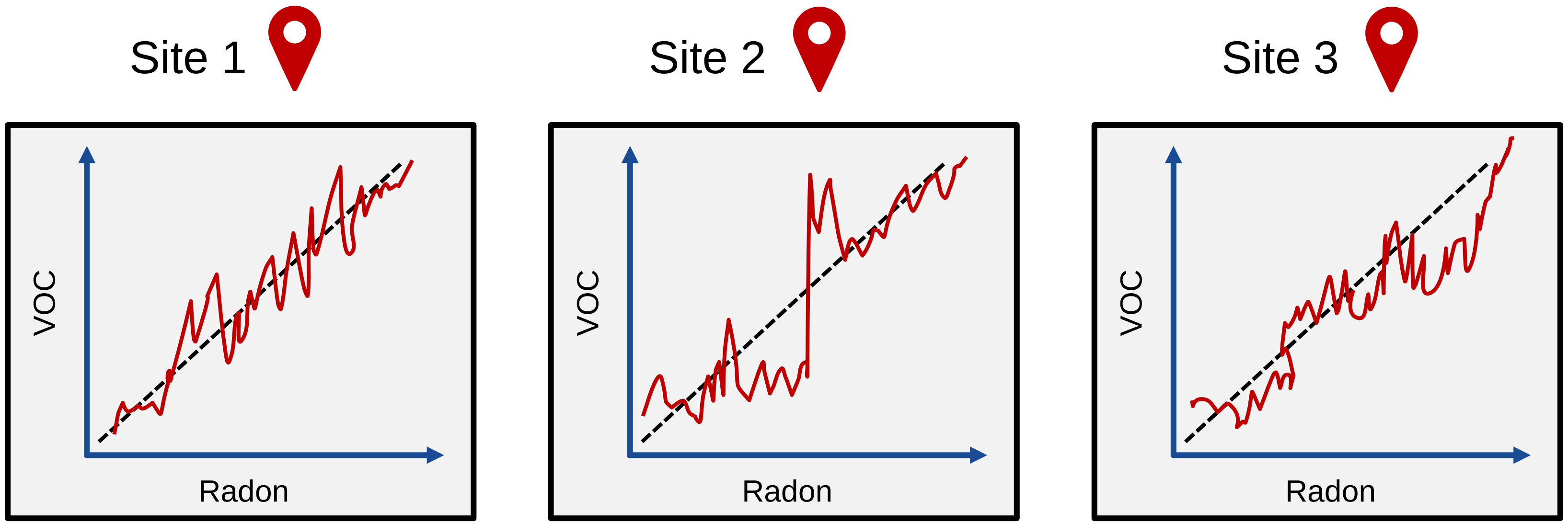
- Many observations of VOC concentrations and predictor variables at sequential points in time from three sites across the country



Actual data set: time series data from three sites

Risk of:

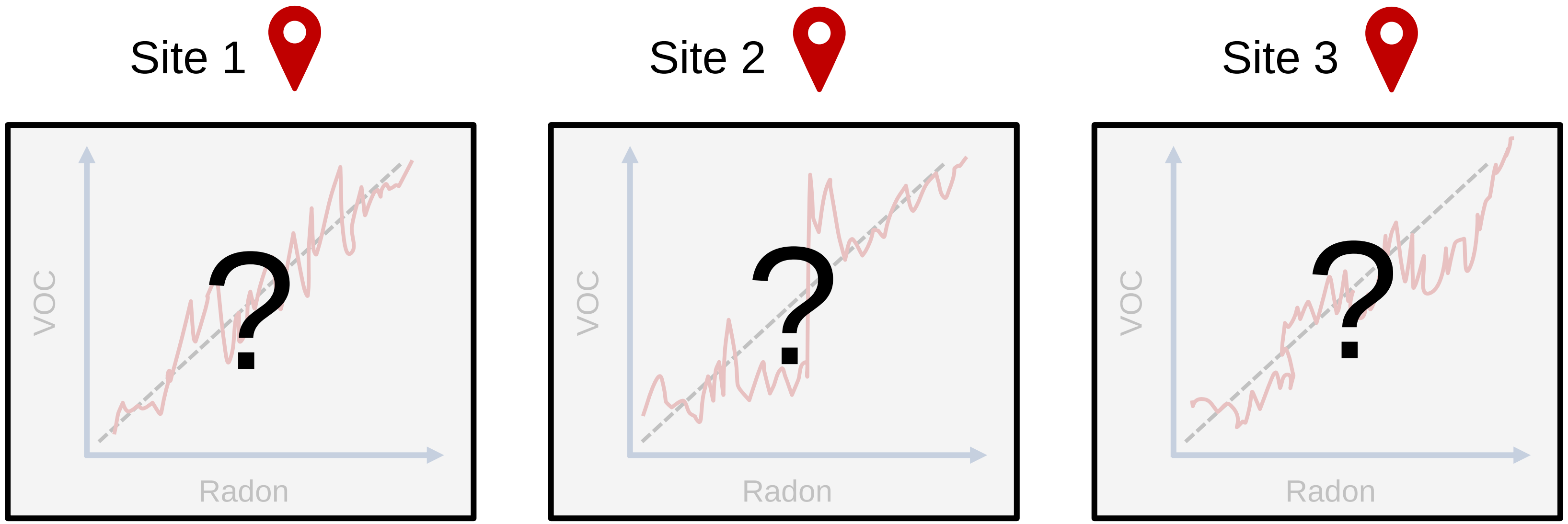
- Autocorrelation and inaccurate estimates
- Spurious regression results
- Systematic bias



Actual data set: time series data from three sites

Risk of:

- Autocorrelation and inaccurate estimates
- Spurious regression results
- Systematic bias



Cannot know the true relationship between the predictor and the outcome variable using traditional linear regression for time series data.

Dynamic Time Series Regression

Refers to the "dynamic" (non-static) nature of time series data where past observations influence current and future values

Regression approach

- **Outcome variable:** Indoor air VOC concentration (C, μg/m³)
- **Predictor variables:**
  - Indoor air radon concentration (R, pCi/L)
  - Differential temperature (ΔT, °C)
  - Differential pressure (ΔP, Pa)

Traditional linear regression:

$$\mathrm{C}_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathrm{R}_t + \boldsymbol{\beta_2}\Delta T_t + \boldsymbol{\beta_3}\Delta P_t + \varepsilon_t$$

Dynamic time series regression:

$$\mathrm{C}_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathrm{R}_t + \boldsymbol{\beta_2}\Delta T_t + \boldsymbol{\beta_3}\Delta P_t + \eta_t$$

"ARIMA errors"

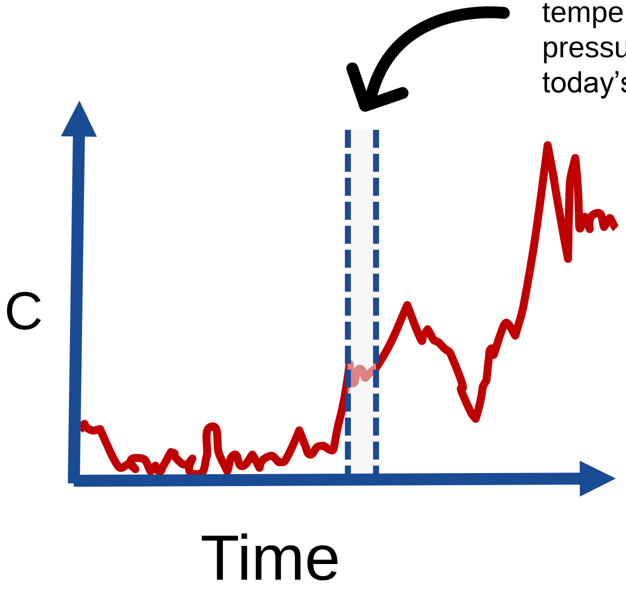
Allows the residuals of the regression to follow an autoregressive integrated moving average (ARIMA) regression

Dynamic Time Series Regression

Traditional linear regression:

$$C_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} R_t + \boldsymbol{\beta_2} \Delta T_t + \boldsymbol{\beta_3} \Delta P_t + \varepsilon_t$$

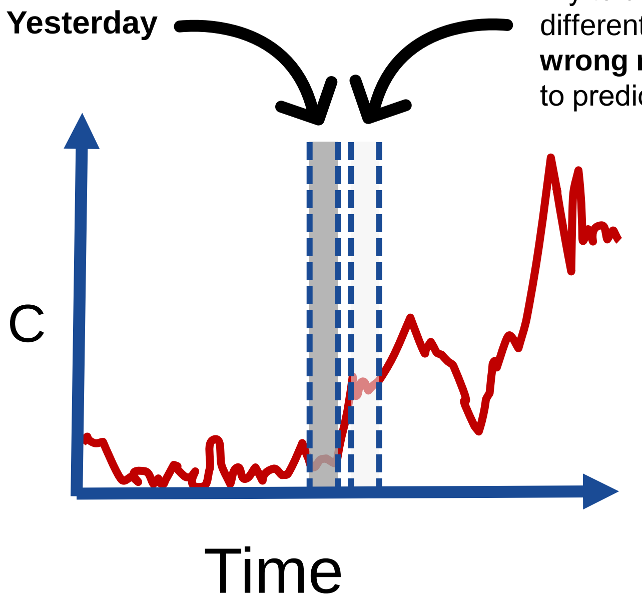
Today
Try to use todays' differential temperature, differential pressure, and radon to predict today's VOC concentration.



Dynamic time series regression:

$$C_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} R_t + \boldsymbol{\beta_2} \Delta T_t + \boldsymbol{\beta_3} \Delta P_t + \eta_t$$

Today
Try to use todays' differential temperature, differential pressure, and radon **and how wrong my prediction was from yesterday** to predict today's VOC concentration.



Allows the regression to "control" for the temporal nature of the data to accurately estimate the true relationship between the predictor variables and the outcome variable.

Data Wrangling and Preprocessing

Virginia Site A

Sun Devil Manor

Indianapolis 422

Data Summary

SITES | **SAMPLE LOCATIONS** | **TIME PERIODS** | **AVERAGING TIMES**

Virginia Site A

Supply Room → 5/2017 – 1/2021

Women’s Restroom → 11/2019 – 1/2021

Office → 7/2019 – 1/2020

Sun Devil Manor

Ground level → 1/2011 – 6/2012

Indianapolis 422

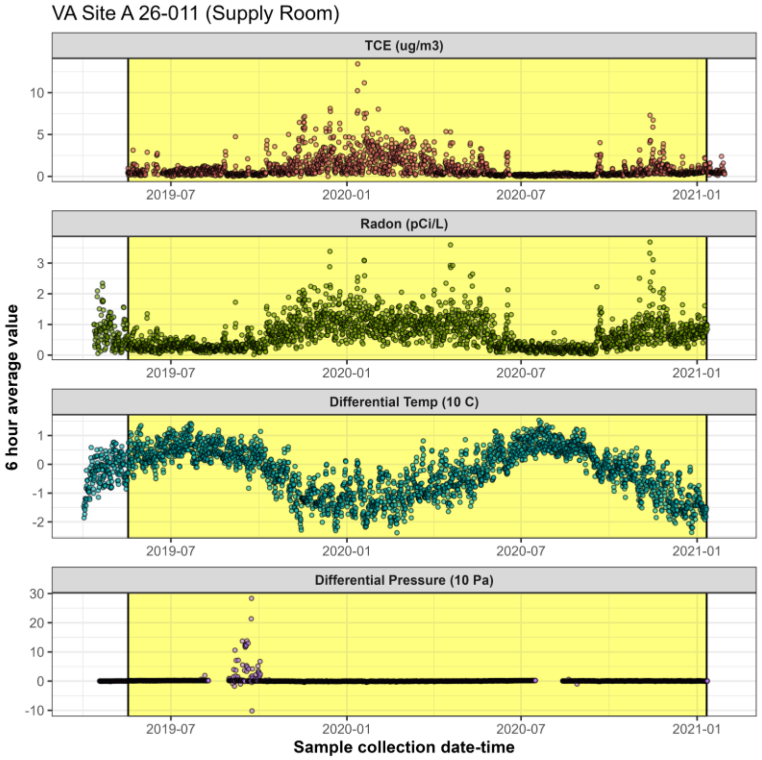
Basement → 8/2011 – 10/2011; 12/2011 – 2/2012

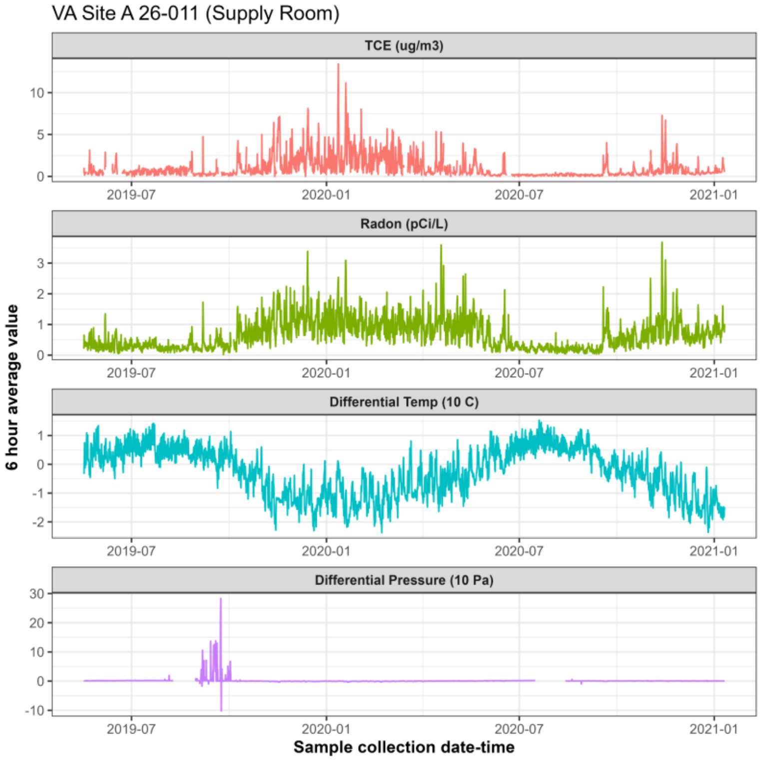
First Floor → 8/2011 – 10/2011; 12/2011 – 2/2012

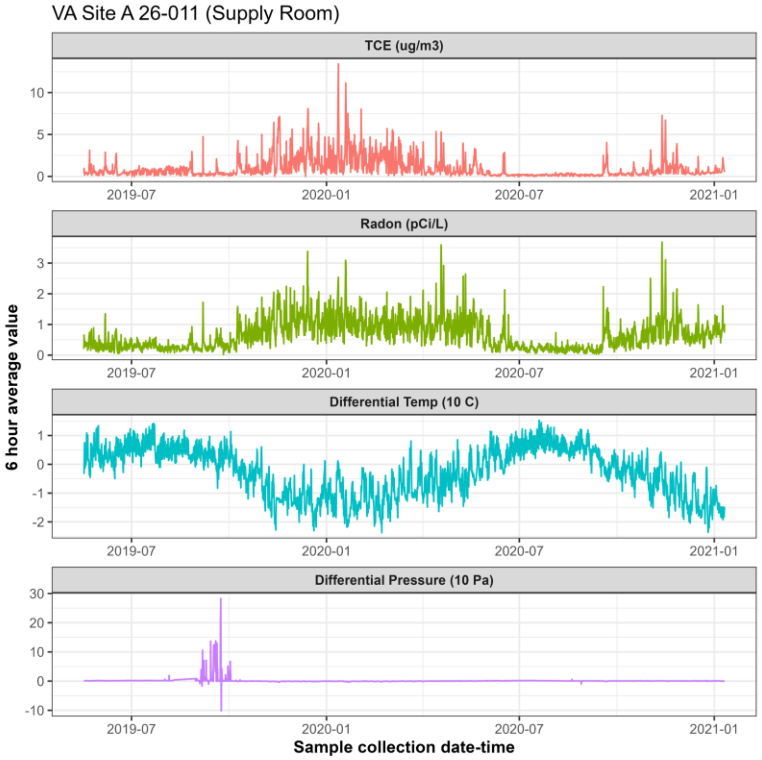
- 6 hour
- 24 hour
- Weekly

Virginia Site A: **Supply Room**

Identify region of dataset with overlapping regressor and response variables → **tsibble** Create time series object → **fable** Interpolate missing data intervals (auto ARIMA)

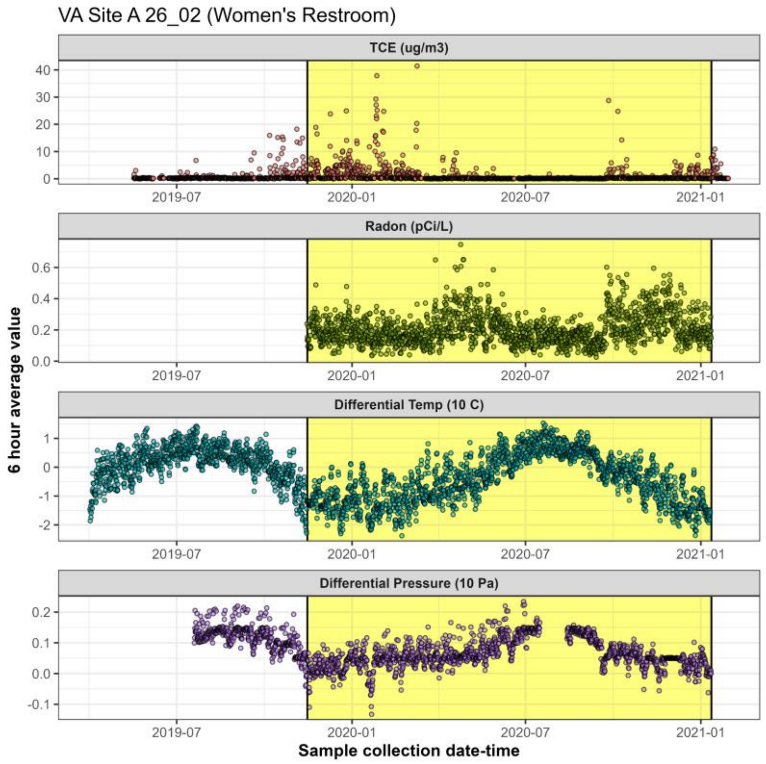


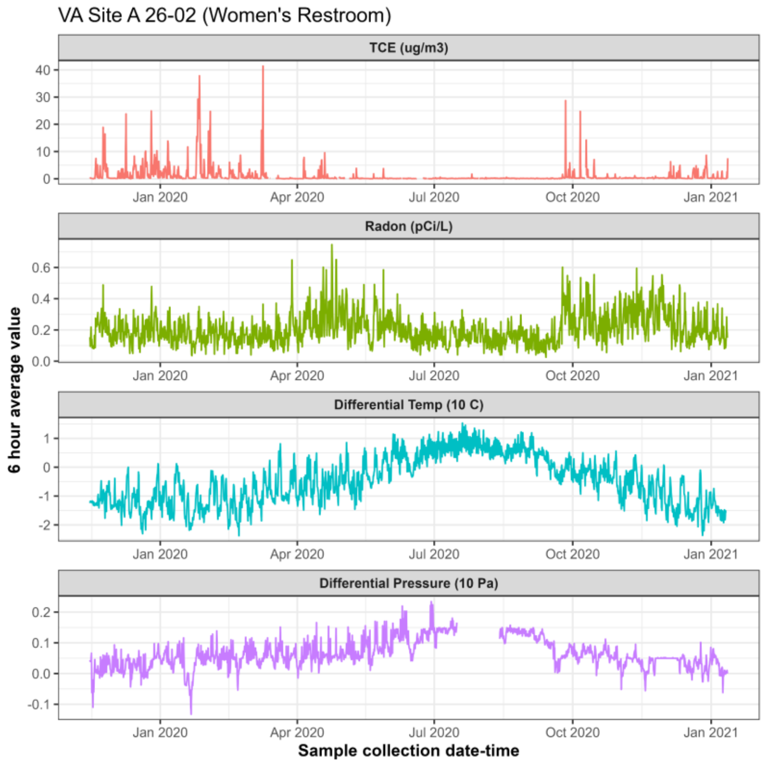


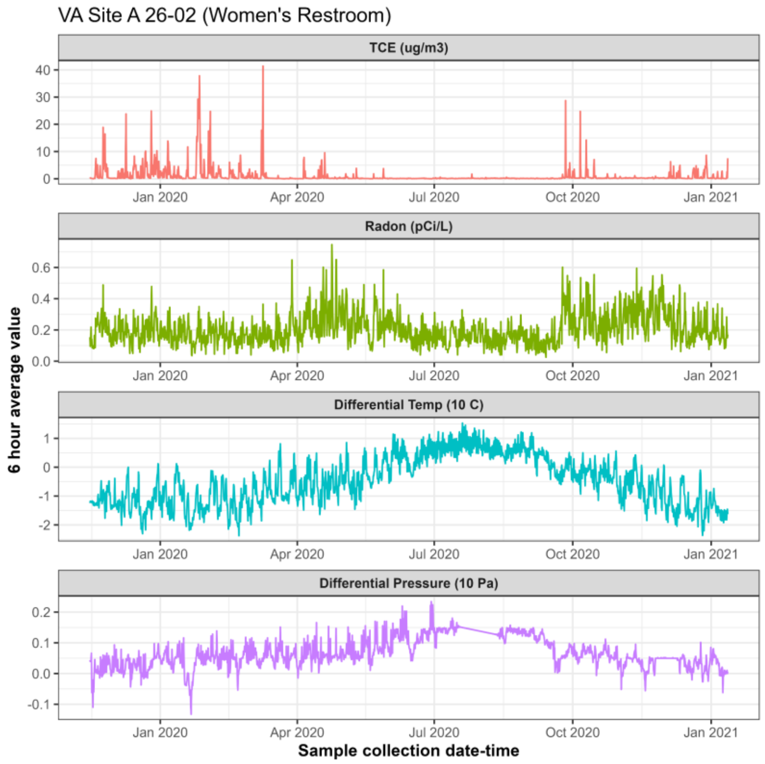


Virginia Site A: **Women's Restroom**

Identify region of dataset with overlapping regressor and response variables → Create time series object → Interpolate missing data intervals (auto ARIMA)

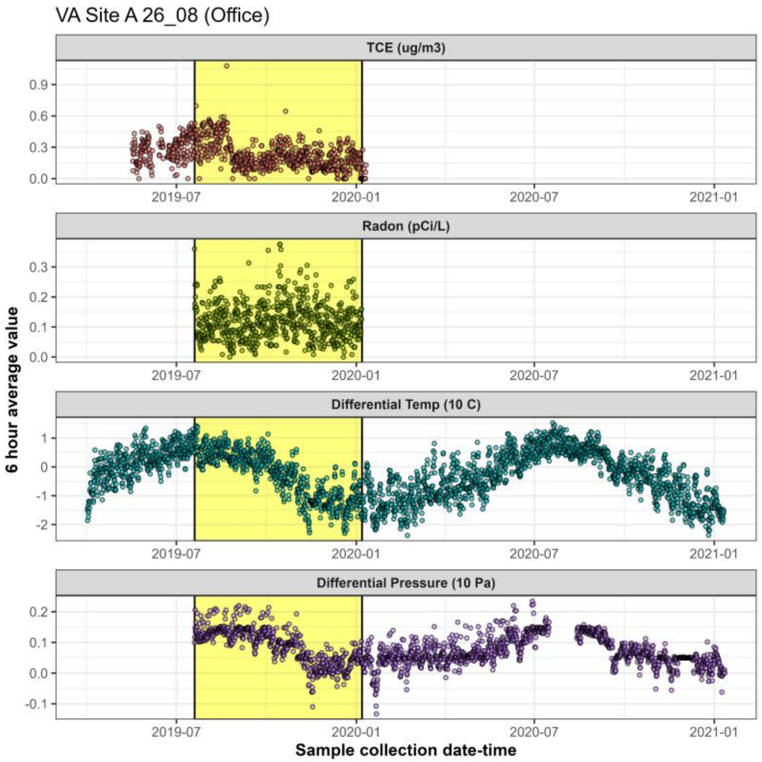


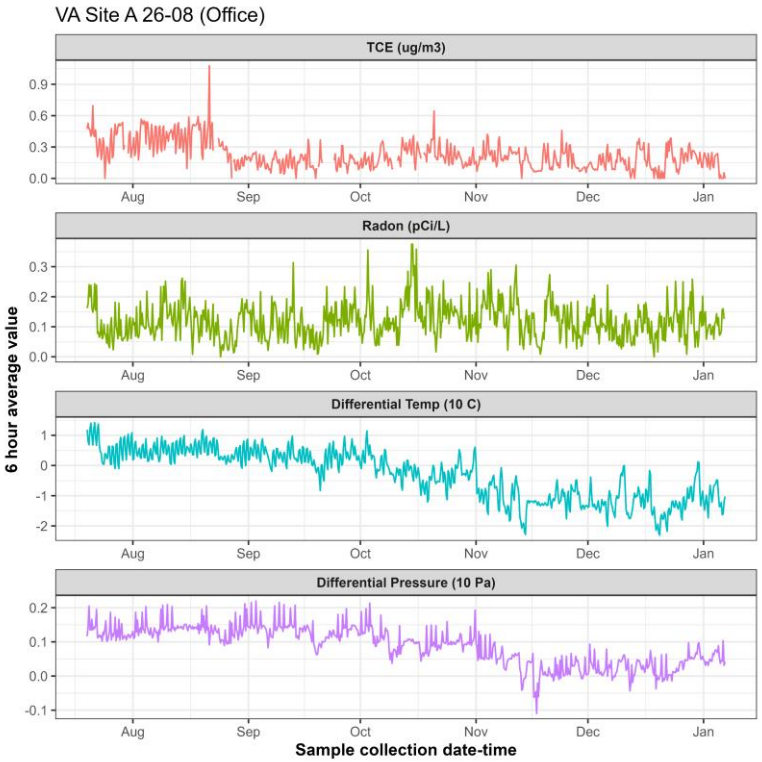


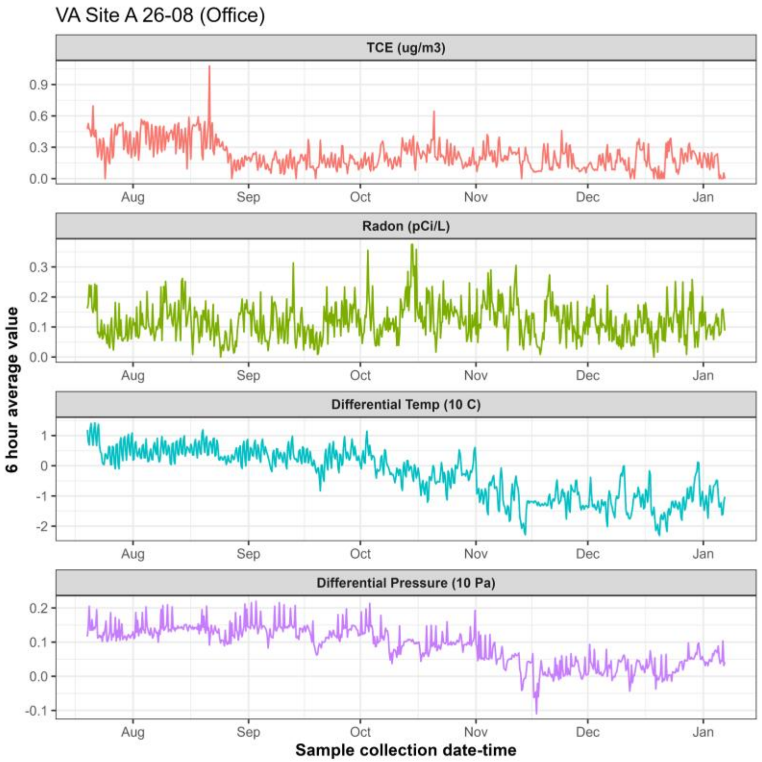


Virginia Site A: **Office**

Identify region of dataset with overlapping regressor and response variables → Create time series object → Interpolate missing data intervals (auto ARIMA)

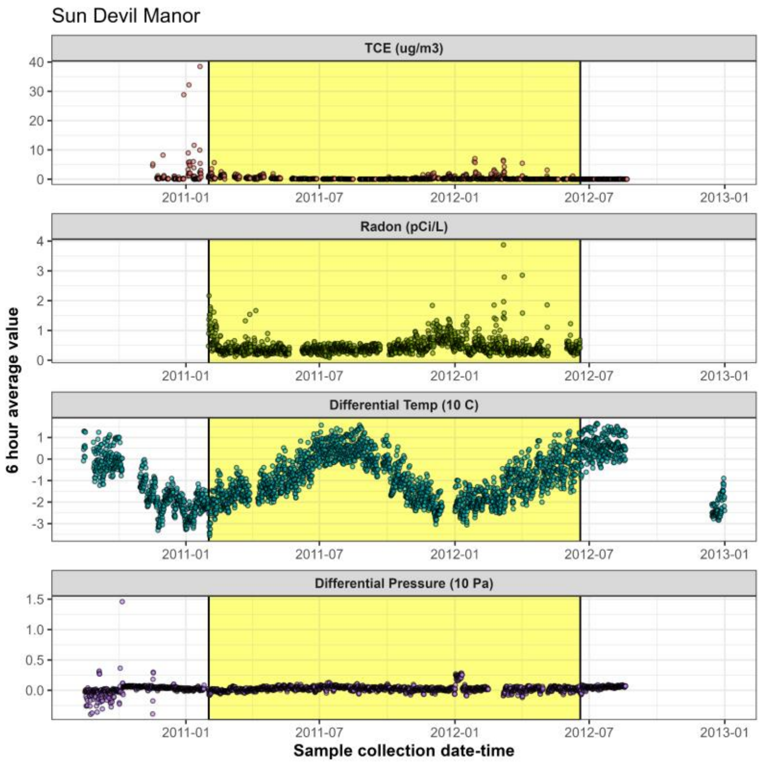


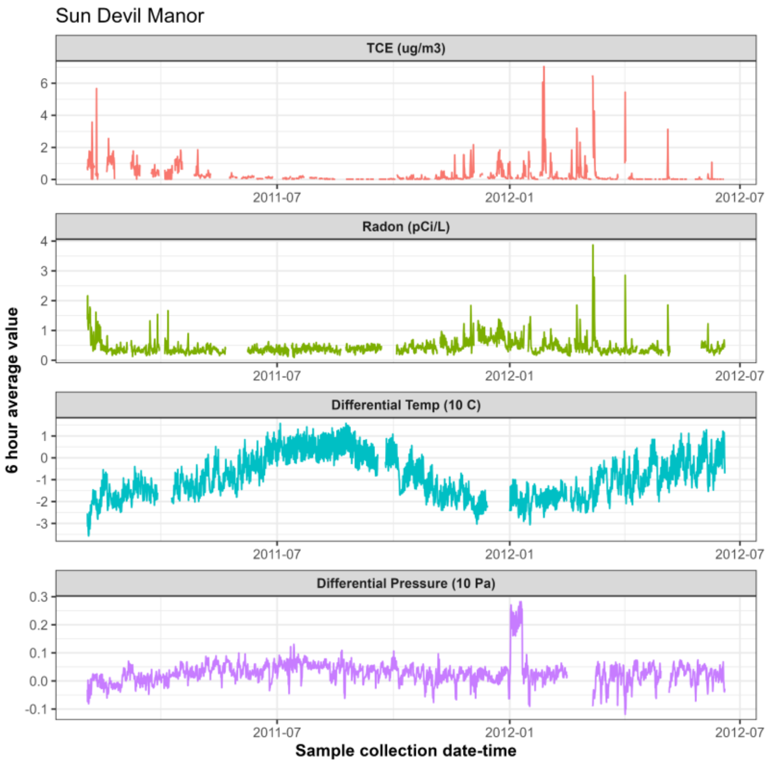


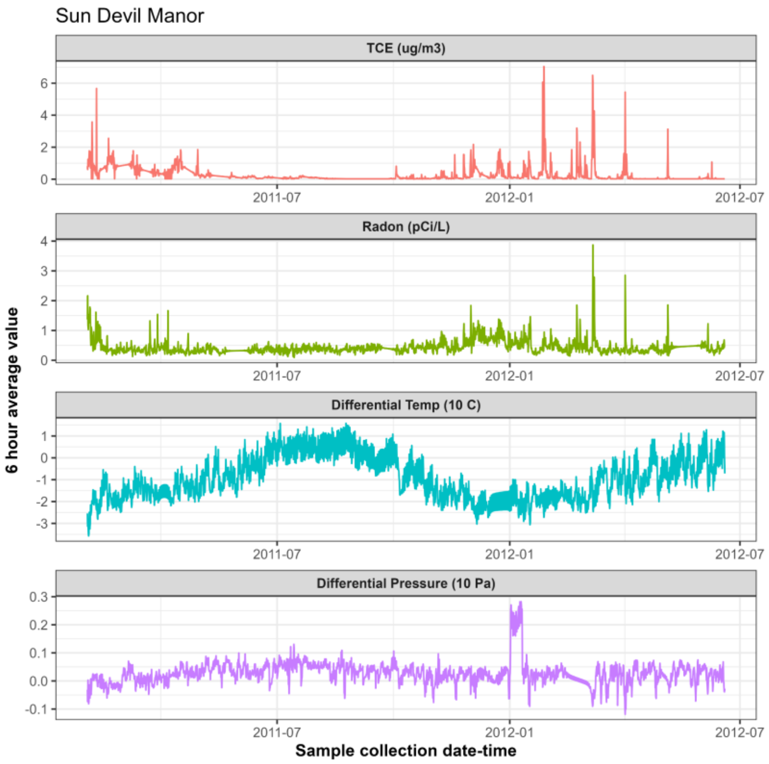


Sun Devil Manor

Identify region of dataset with overlapping regressor and response variables → Create time series object → Interpolate missing data intervals (auto ARIMA)

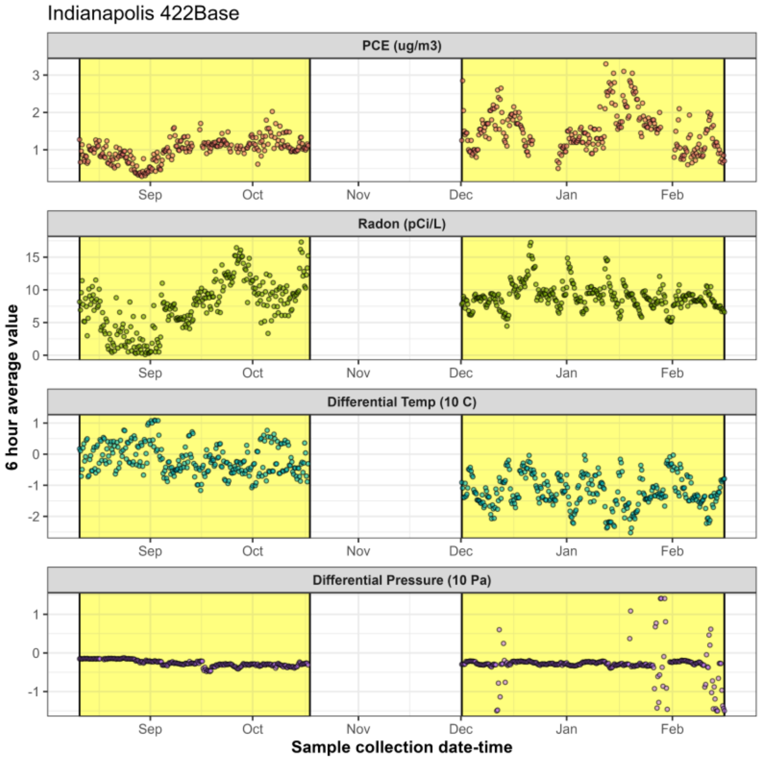


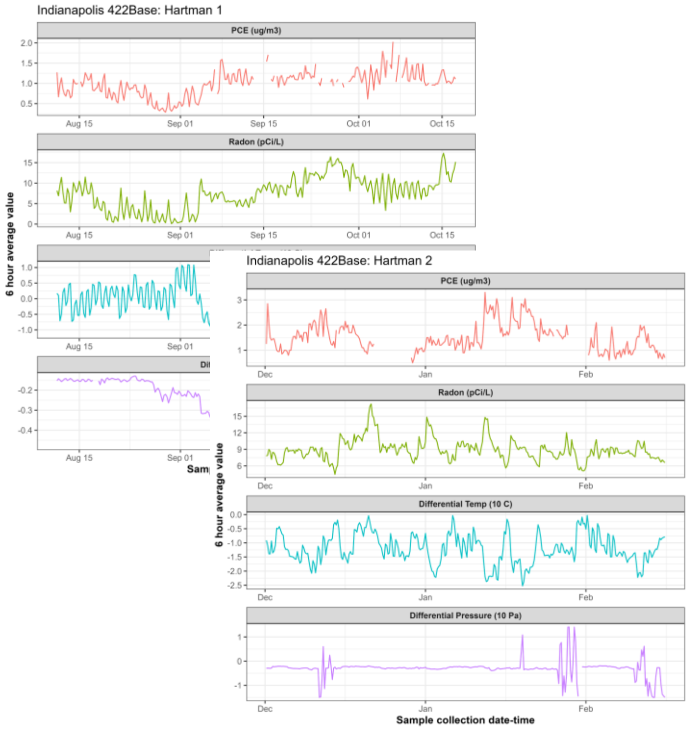




Indianapolis 422 **Basement**

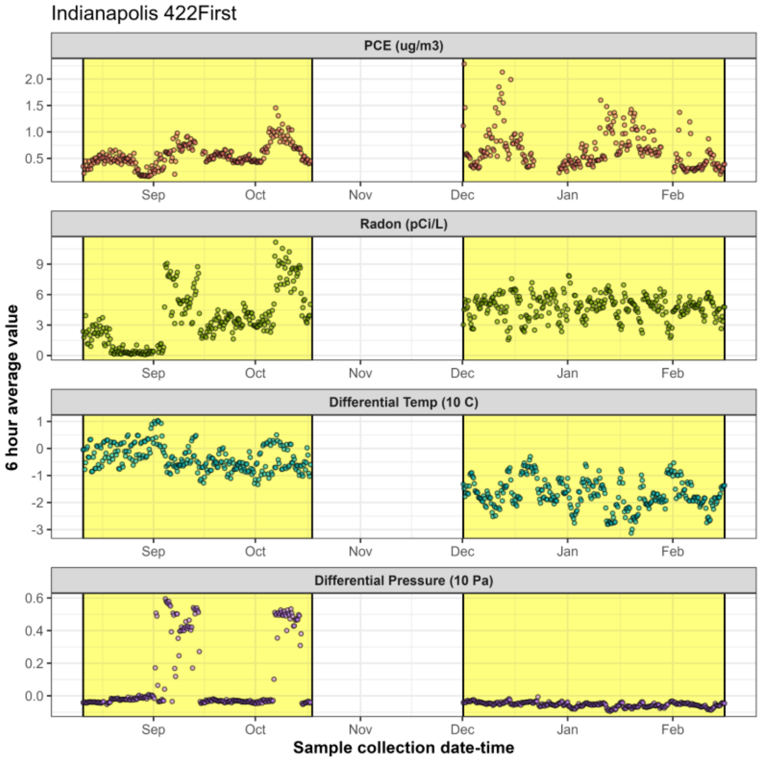
Identify region of dataset with overlapping regressor and response variables → Create time series object → Interpolate missing data intervals (auto ARIMA)

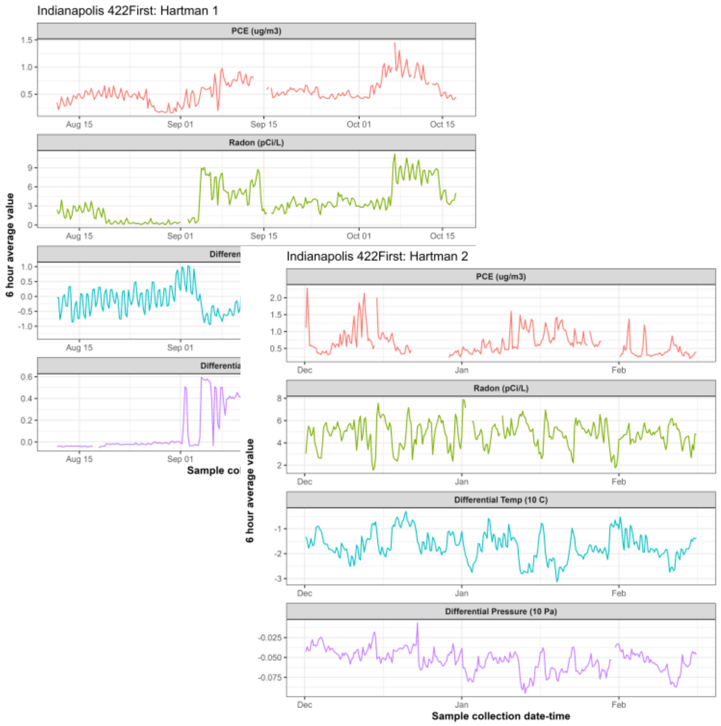


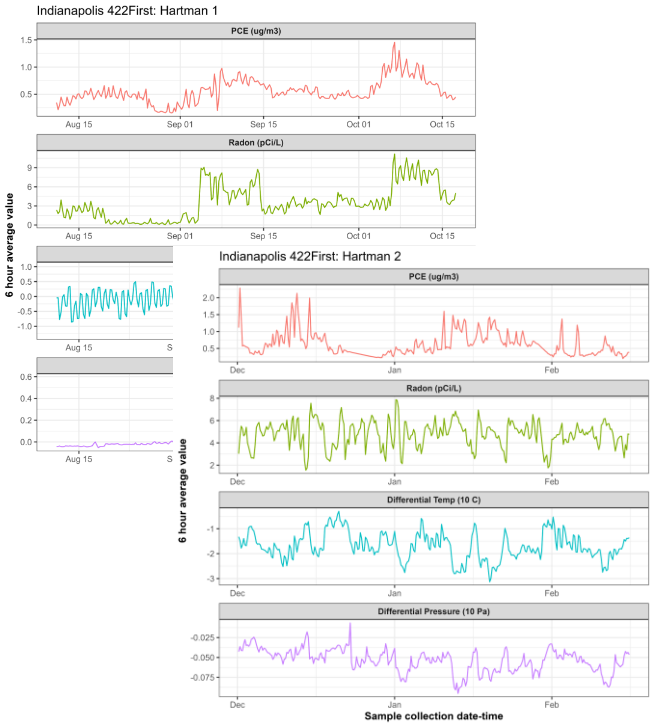


Indianapolis 422 **First Floor**

Identify region of dataset with overlapping regressor and response variables → Create time series object → Interpolate missing data intervals (auto ARIMA)

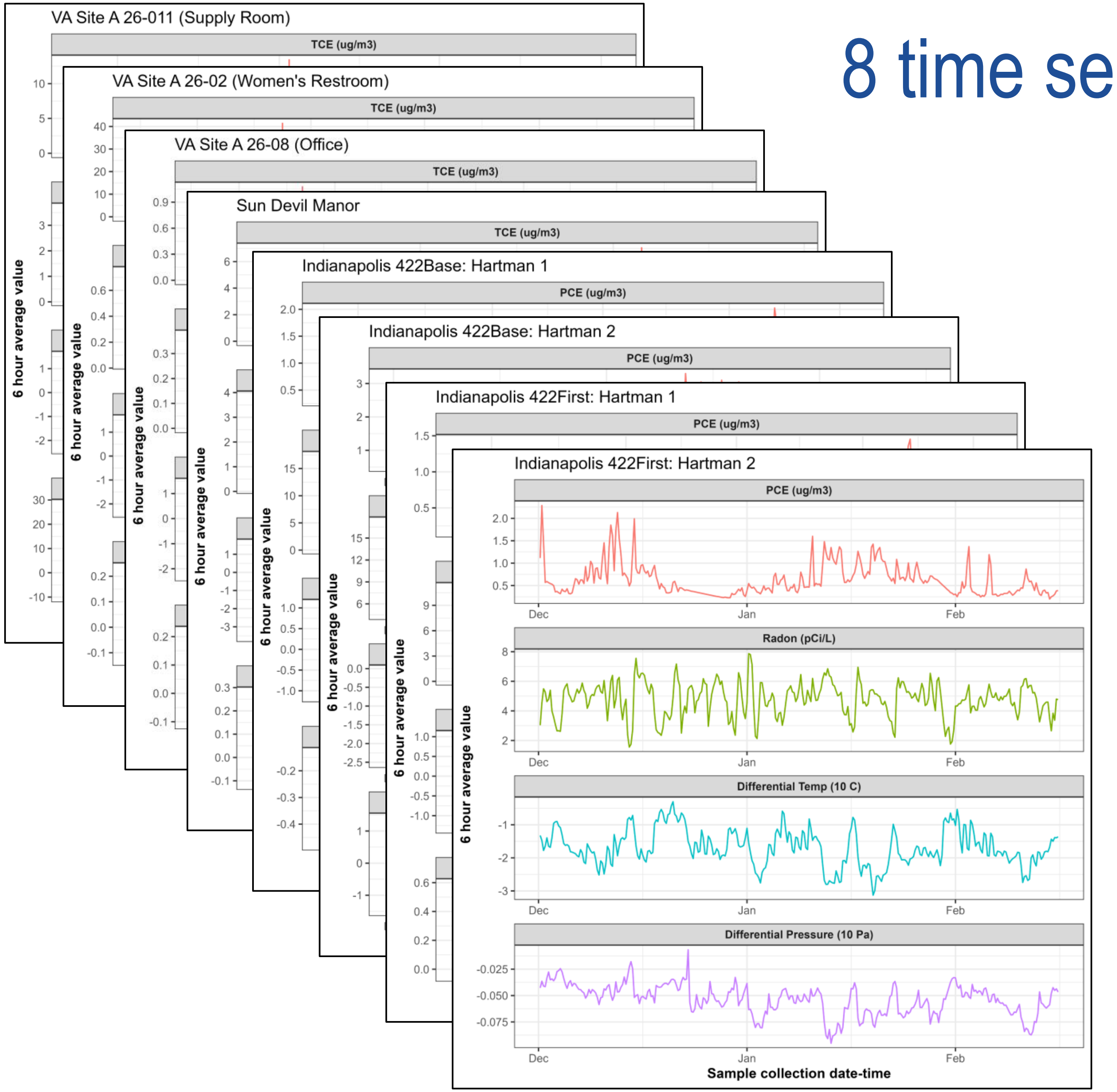






Regression Development

8 time series, 24 multivariate regressions



- 3 **sites**
  - 6 total **sample locations**
    - 8 total **time periods**

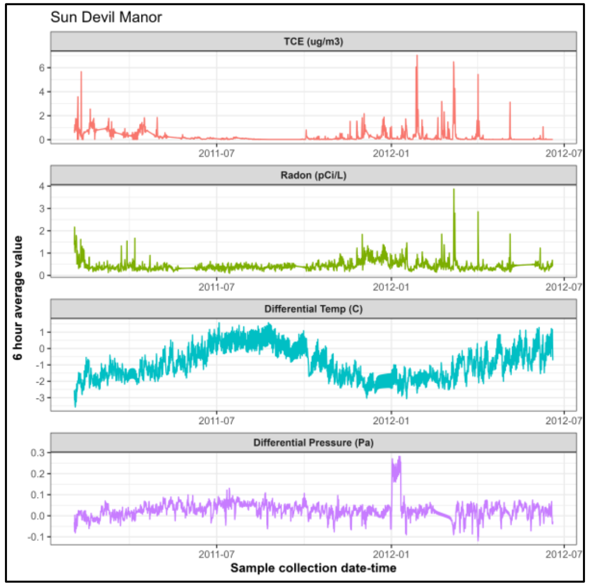
- Each time series modeled using **three different averaging times**
  - 6 hour
  - 24 hour
  - Weekly

- All regression combinations tested:
  - 24 **complete models** with all predictors
  - 72 **“leave-one-out” models**
  - 72 **single variate models**
  - 168 total models

Develop a separate regression for each time series and averaging time

Sun Devil Manor

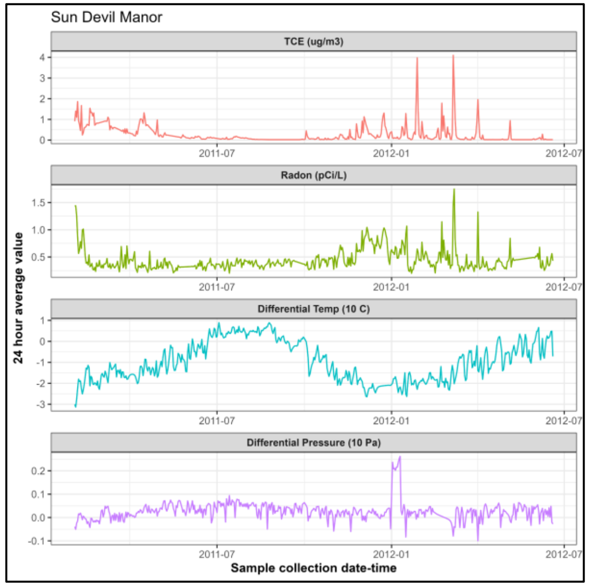
6 hour



Outcome variable

Predictor variables

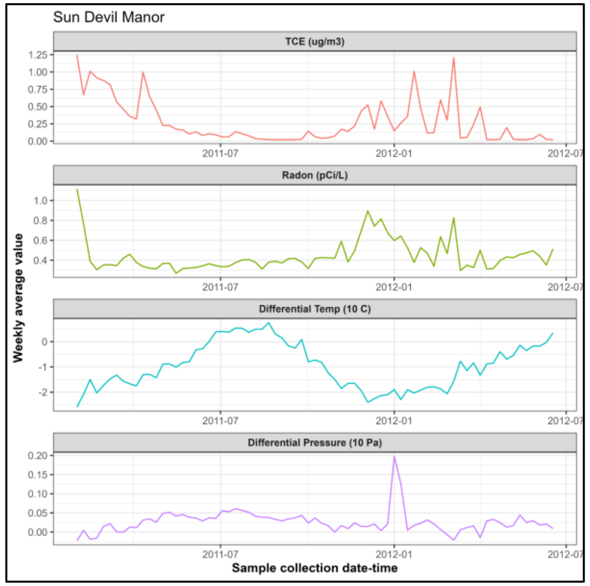
24 hour



Outcome variable

Predictor variables

Weekly



Outcome variable

Predictor variables

$$C_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 R_t + \boldsymbol{\beta}_2 \Delta T_t + \boldsymbol{\beta}_3 \Delta P_t + \eta_t$$

fable

feasts

Results

Virginia Site A

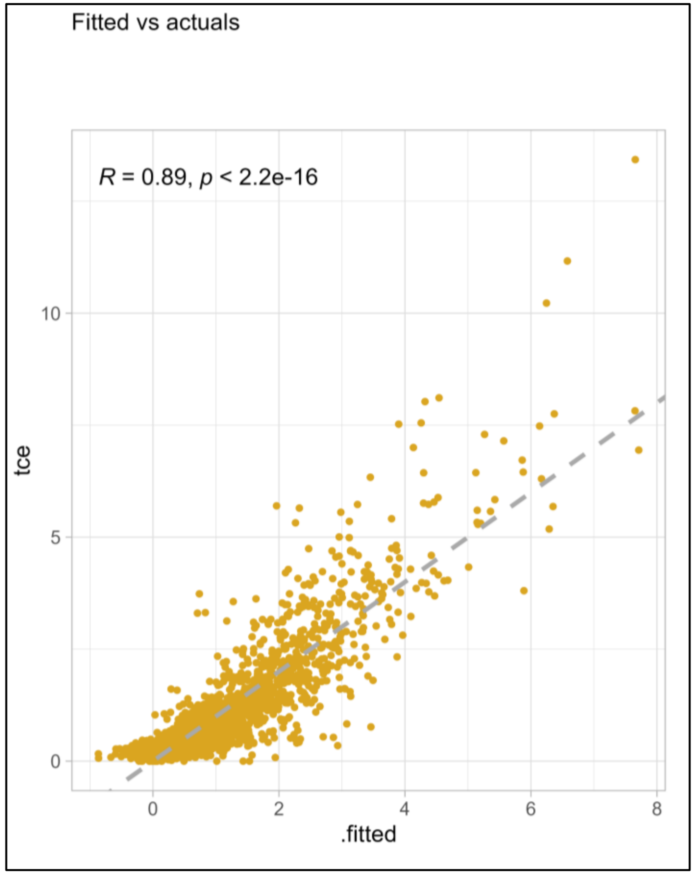
Sun Devil Manor

Indianapolis 422

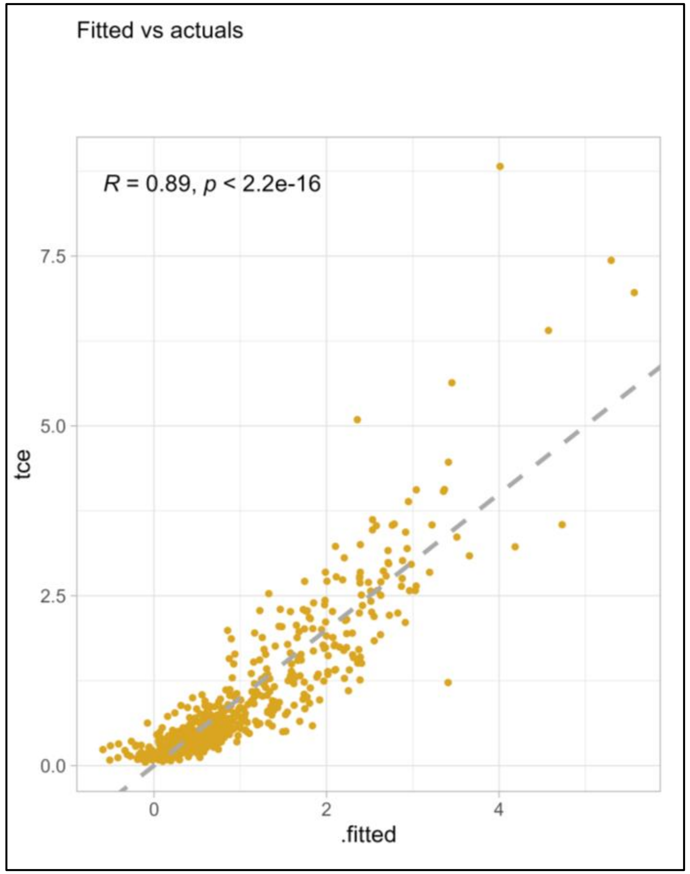
Virginia Site A: **Supply Room**

Complete Model Fit

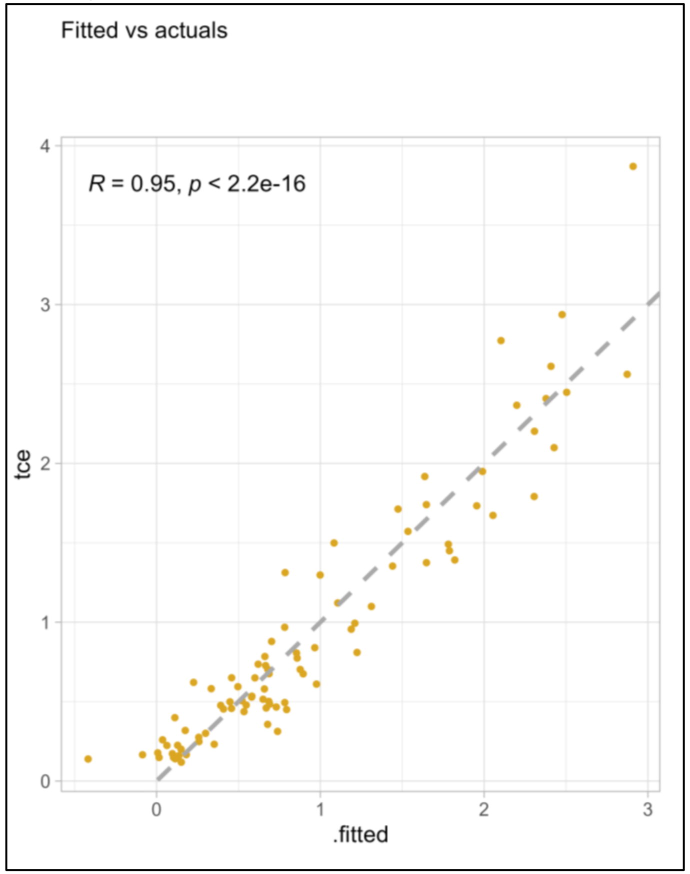
6 hour



24 hour

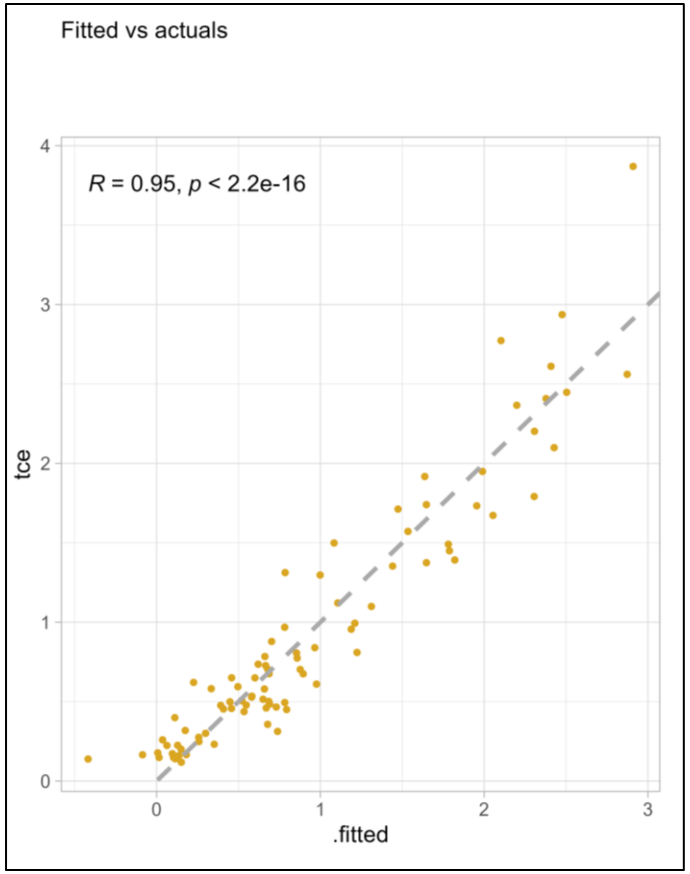


Weekly

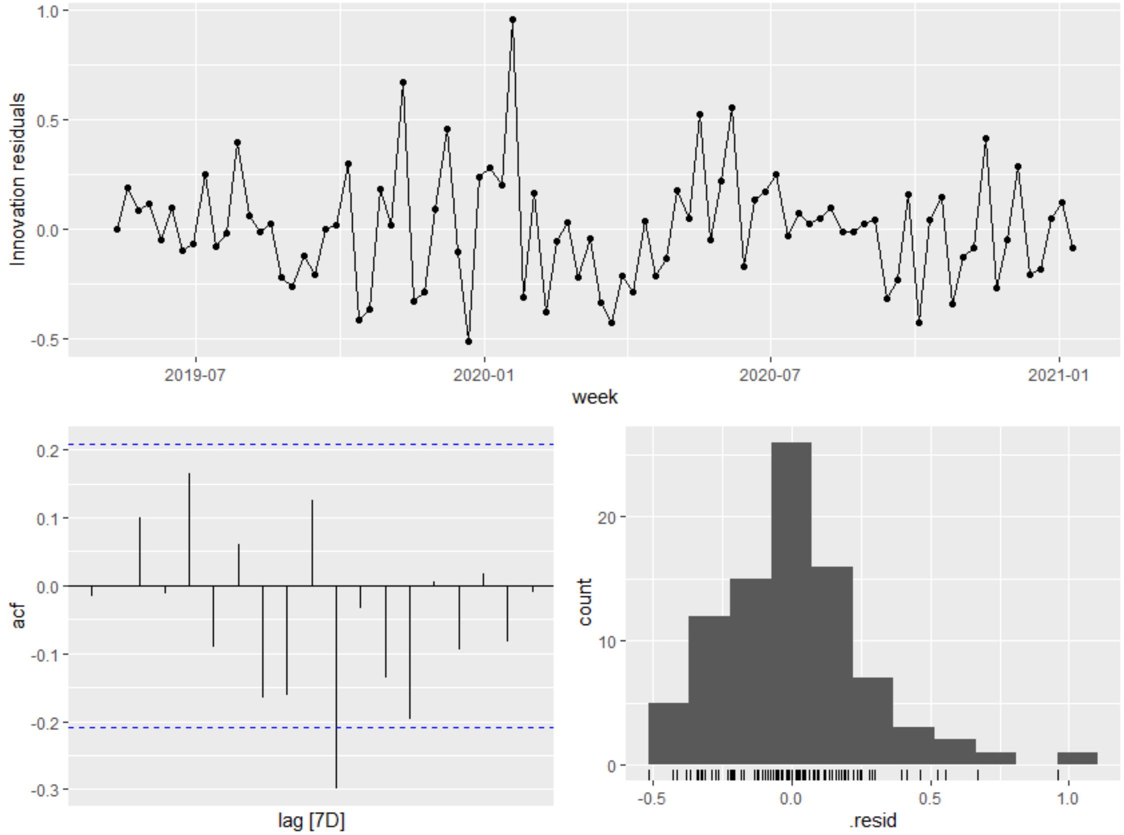


Virginia Site A: **Supply Room**
Complete Regression Diagnostics

Weekly



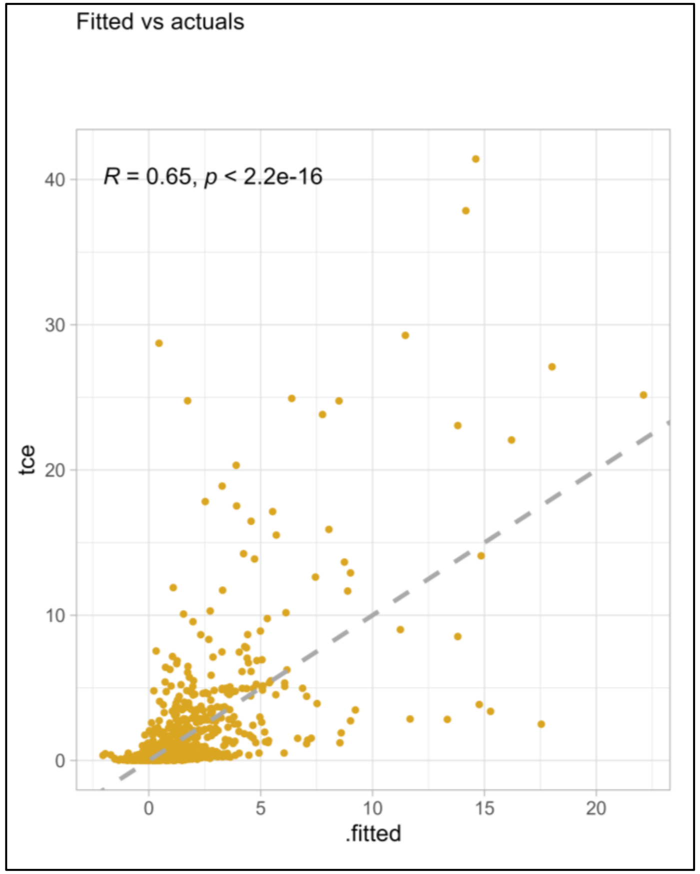
Residuals should resemble a white noise series

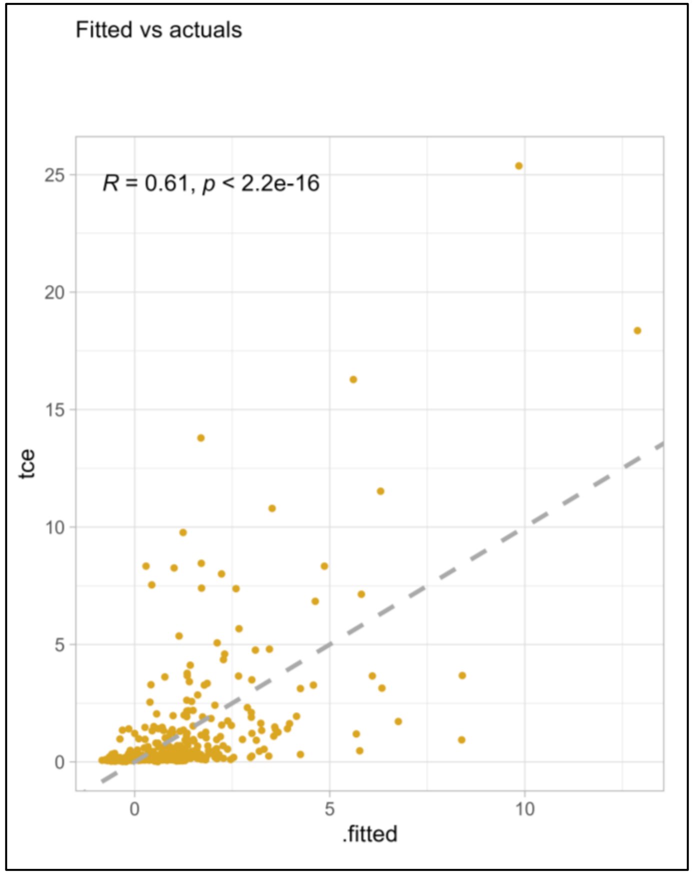


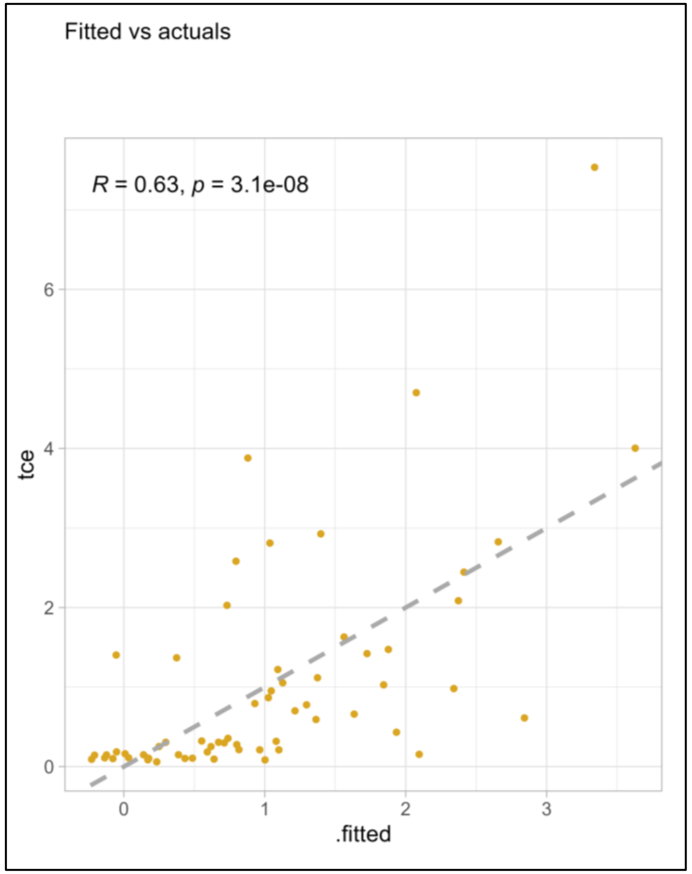
Minimal autocorrelation

Residuals should represent a normal distribution

Virginia Site A: **Women's Restroom**
Complete Model Fit

6 hour


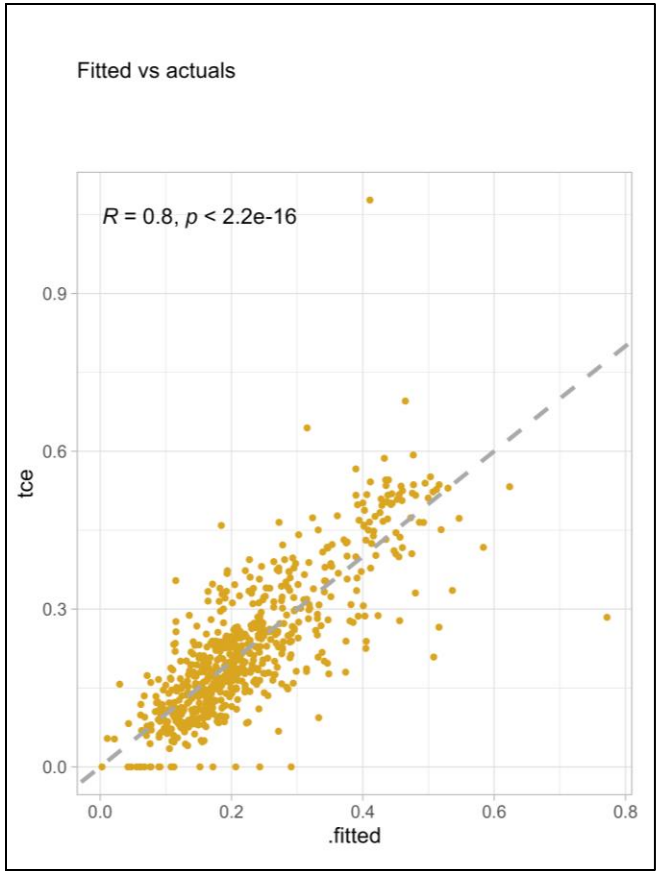
24 hour


Weekly


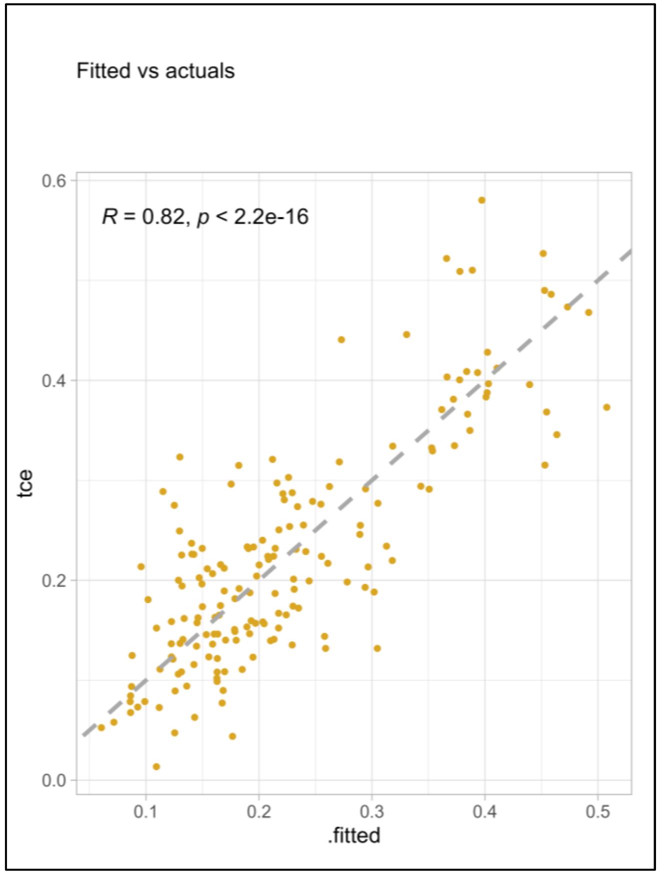
Virginia Site A: **Office**

Complete Model Fit

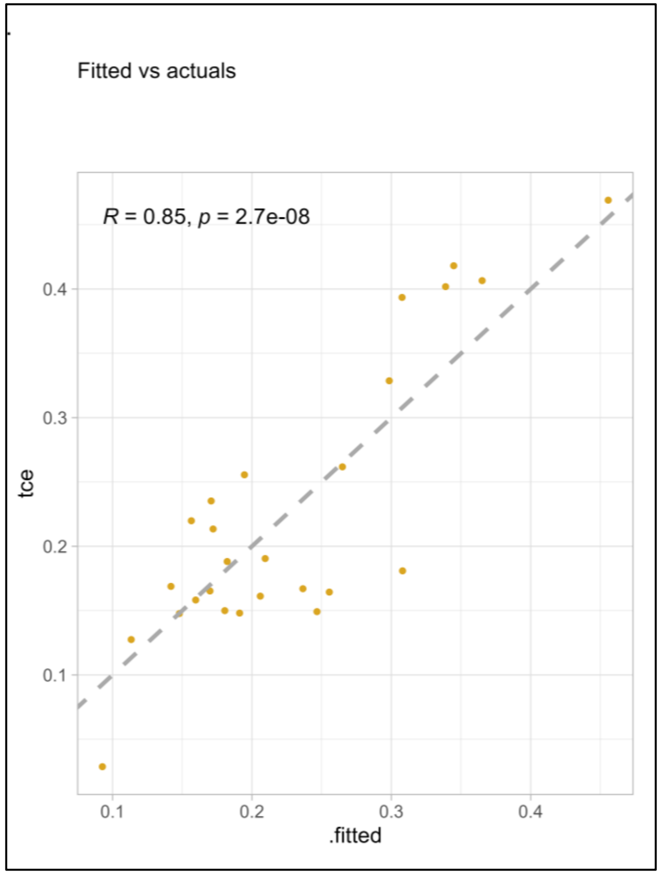
6 hour



24 hour



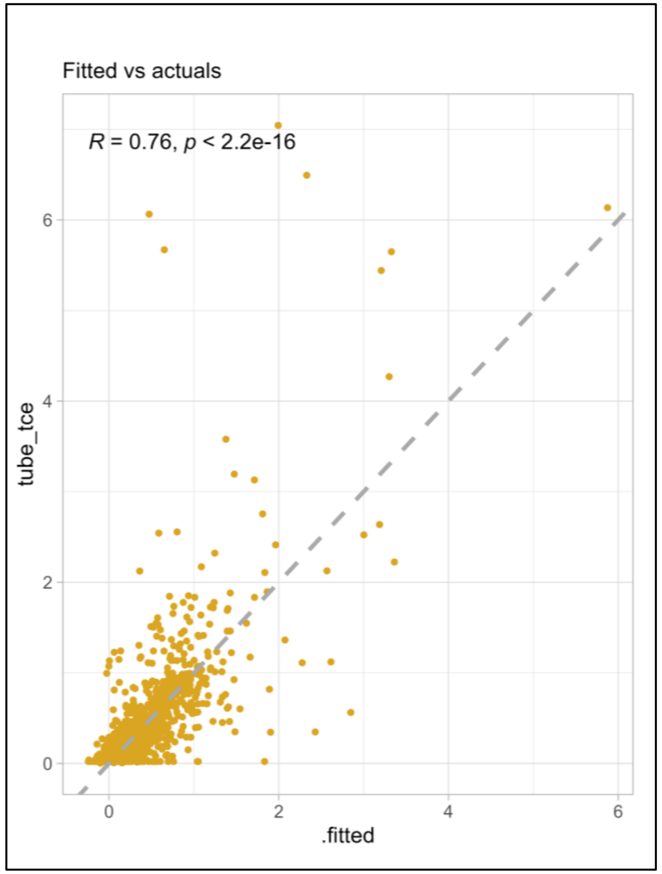
Weekly



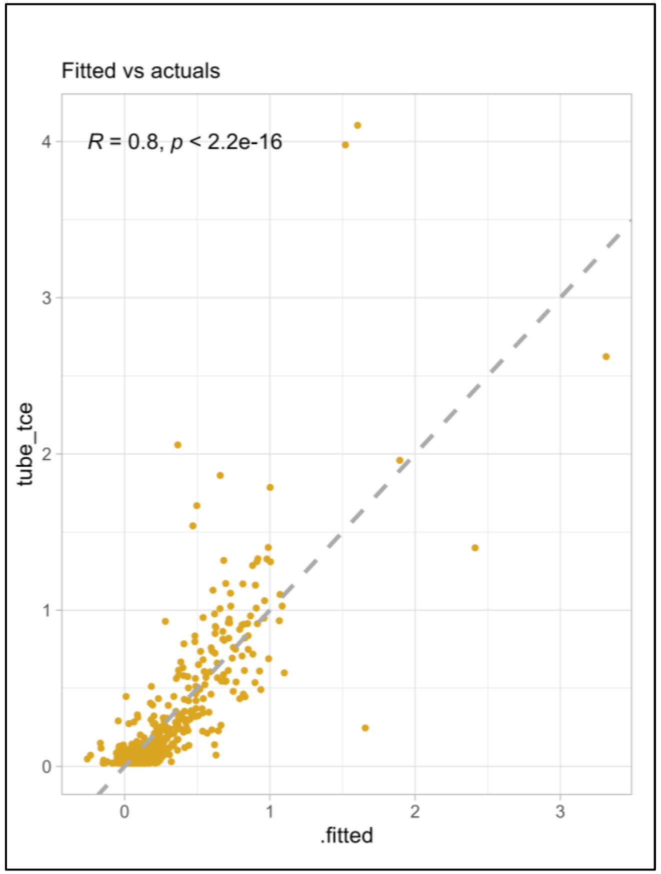
Sun Devil Manor

Complete Model Fit

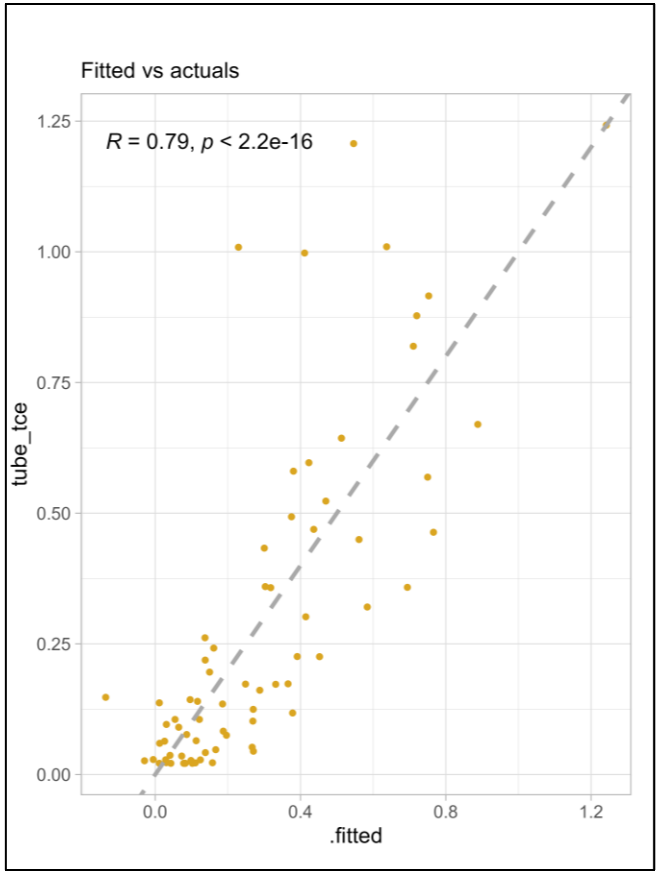
6 hour



24 hour

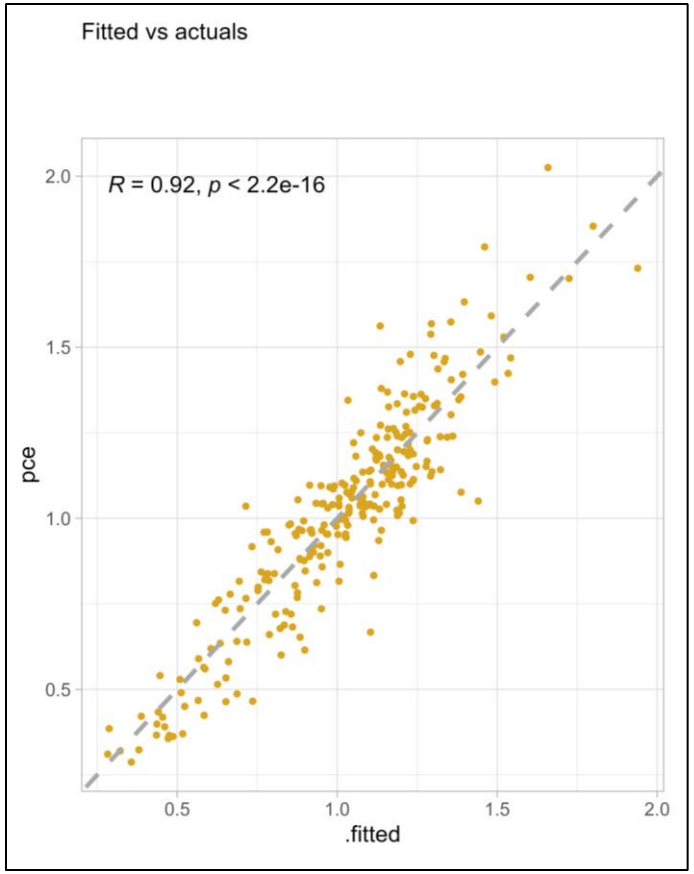


Weekly

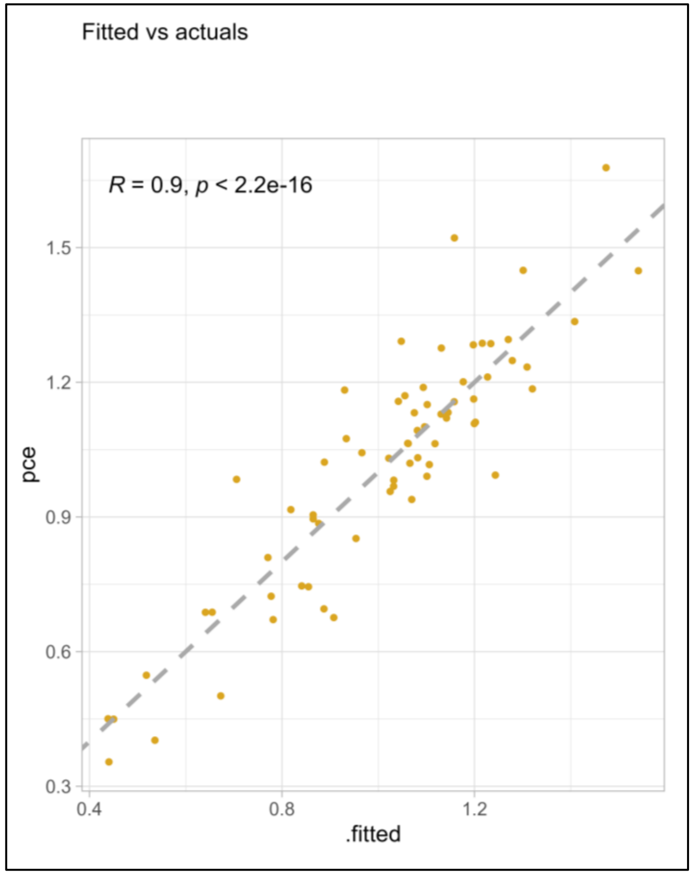


Indianapolis 422 **Basement**: Aug-Oct
Complete Model Fit

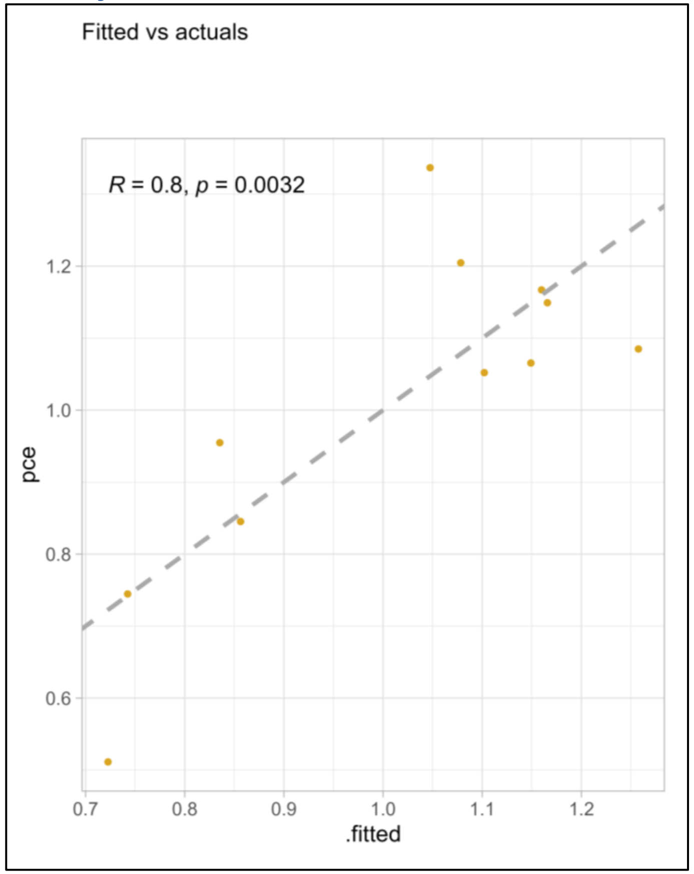
6 hour



24 hour



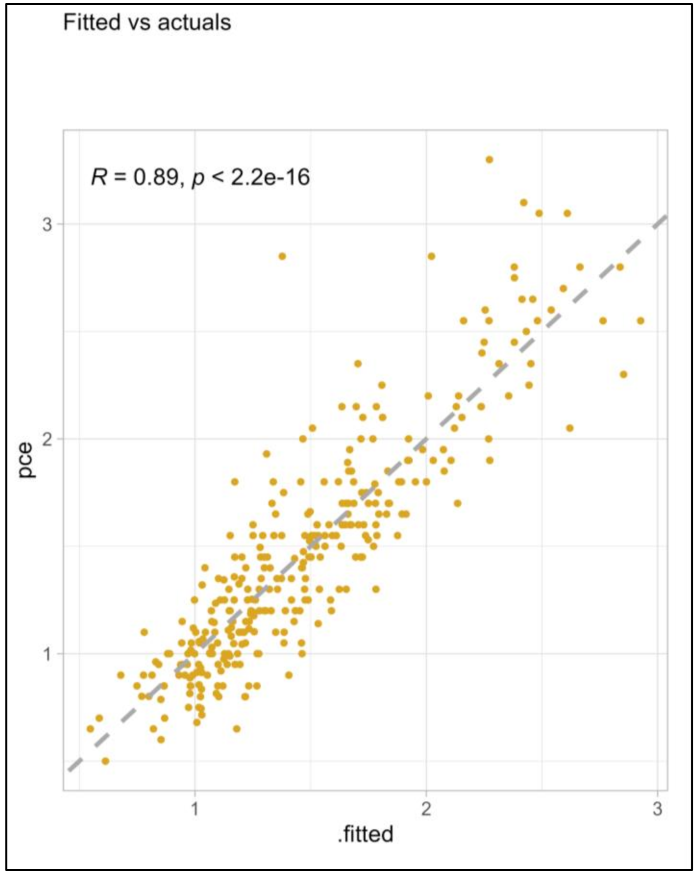
Weekly



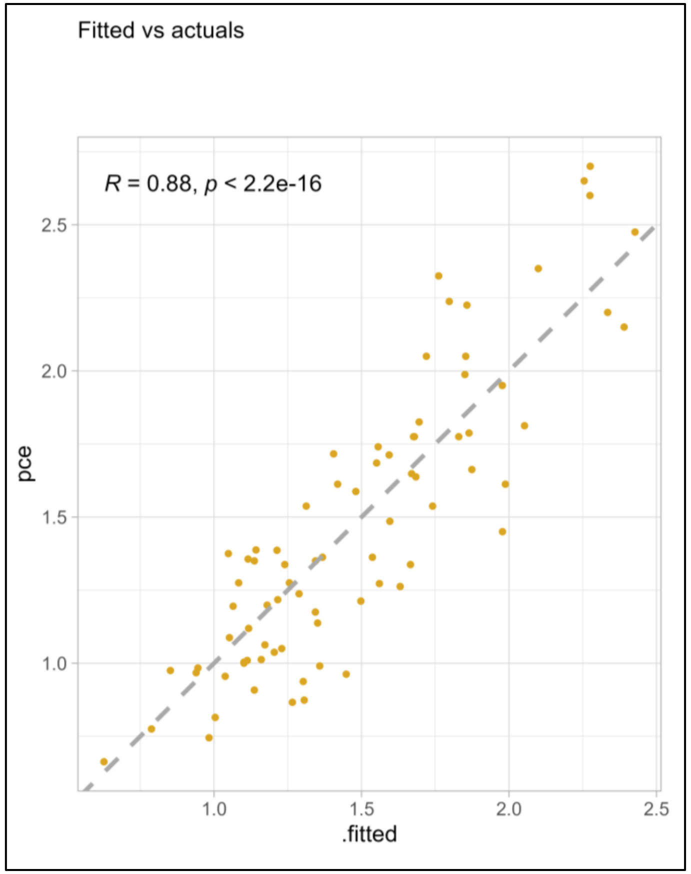
Indianapolis 422 **Basement**: Dec-Feb

Complete Model Fit

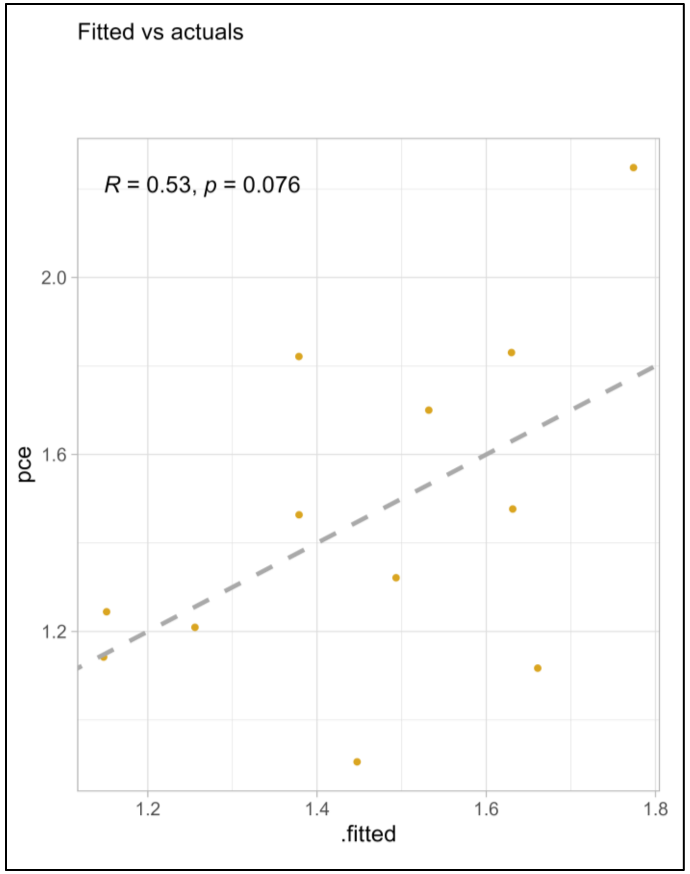
6 hour



24 hour



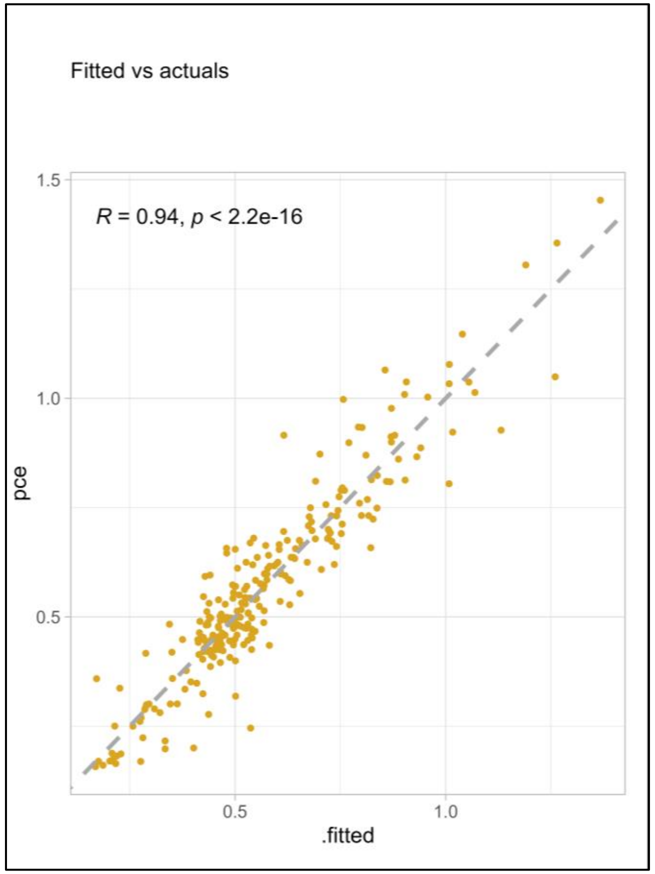
Weekly



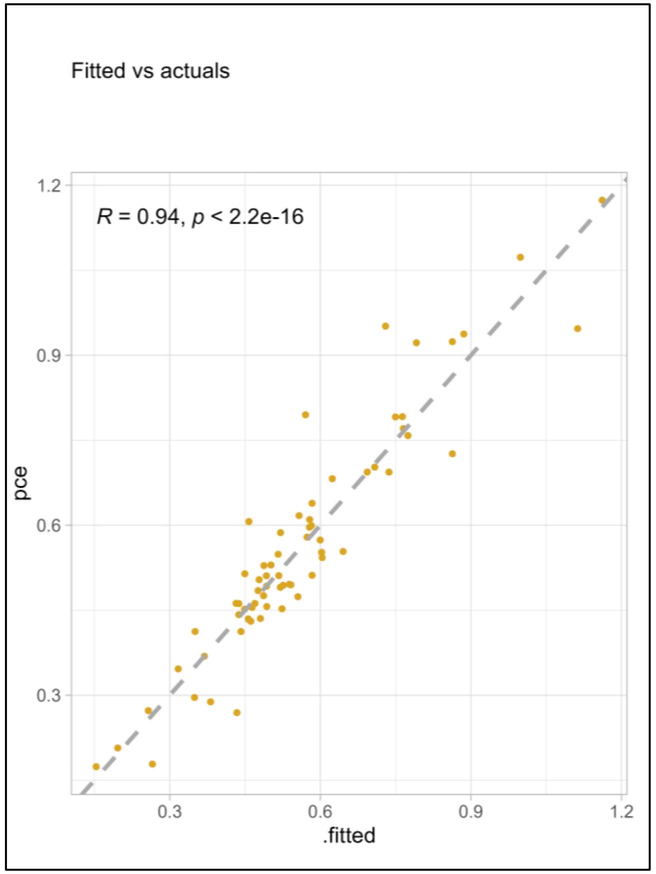
Indianapolis 422 **First Floor**: Aug-Oct

Complete Model Fit

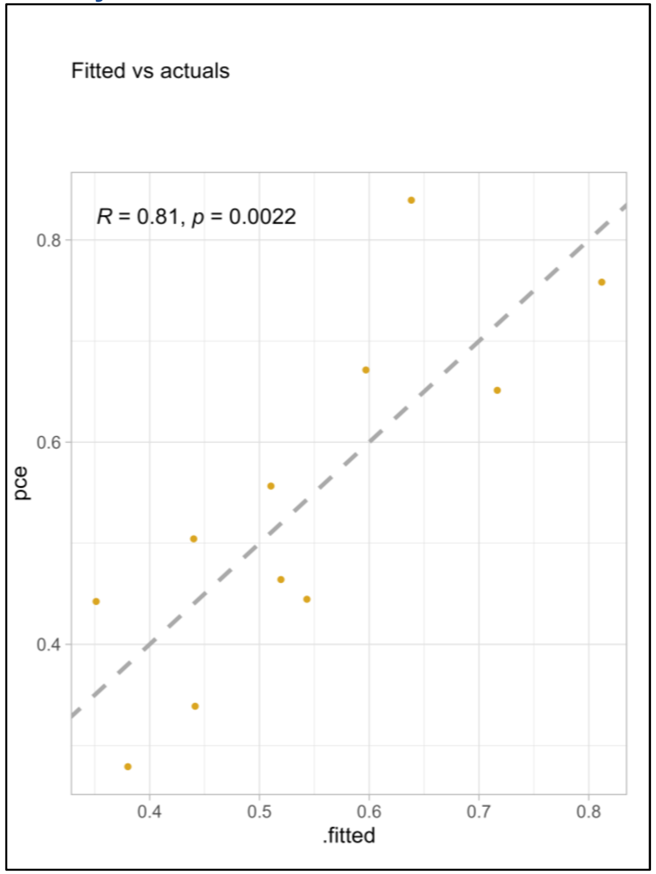
6 hour



24 hour

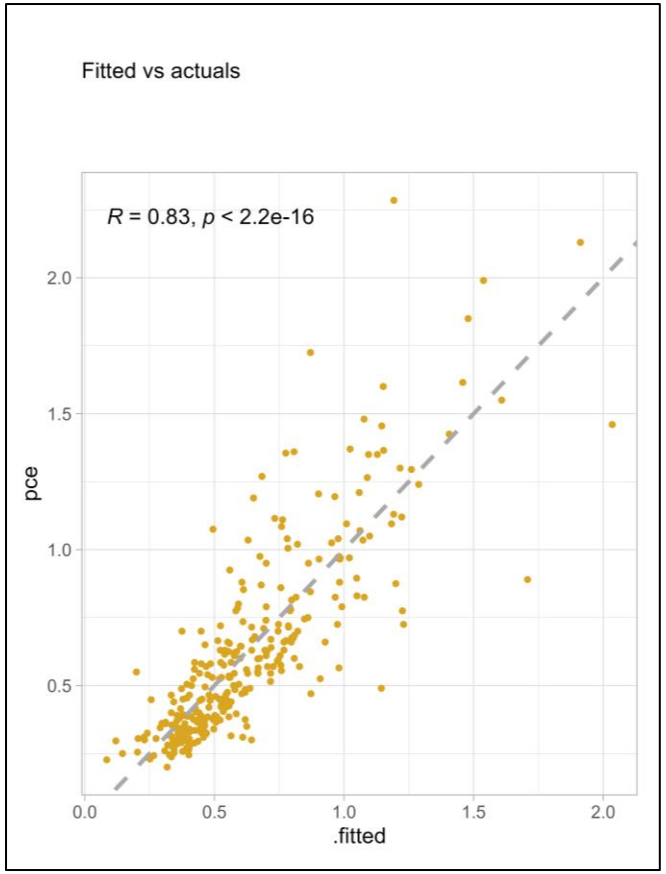


Weekly

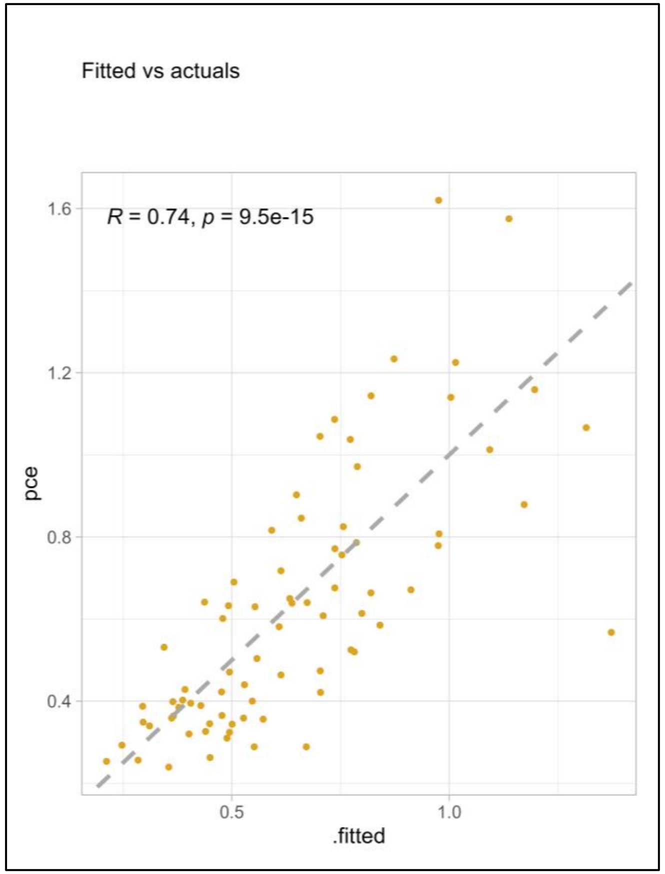


Indianapolis 422 **First Floor**: Dec-Feb
Complete Model Fit

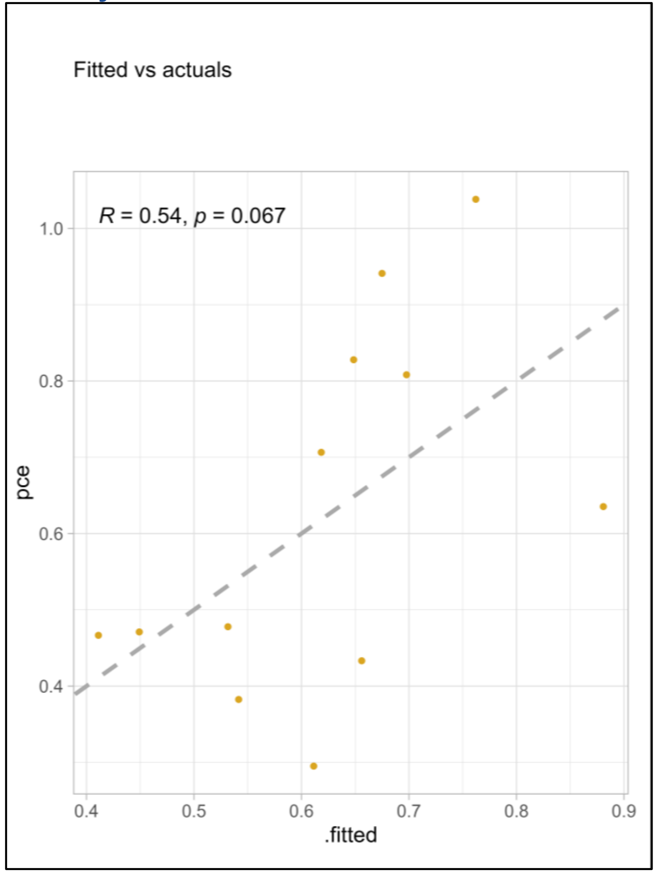
6 hour



24 hour

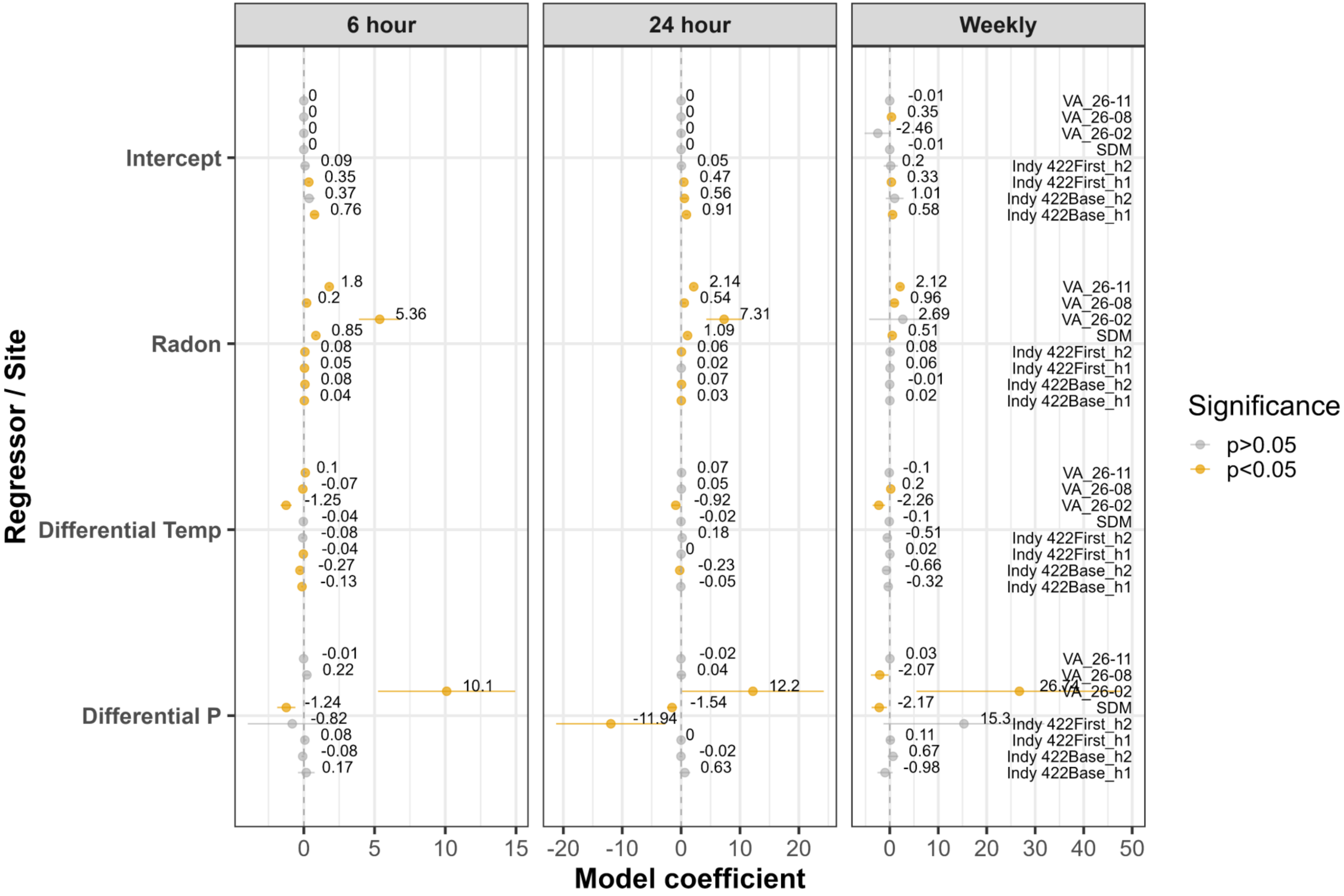


Weekly



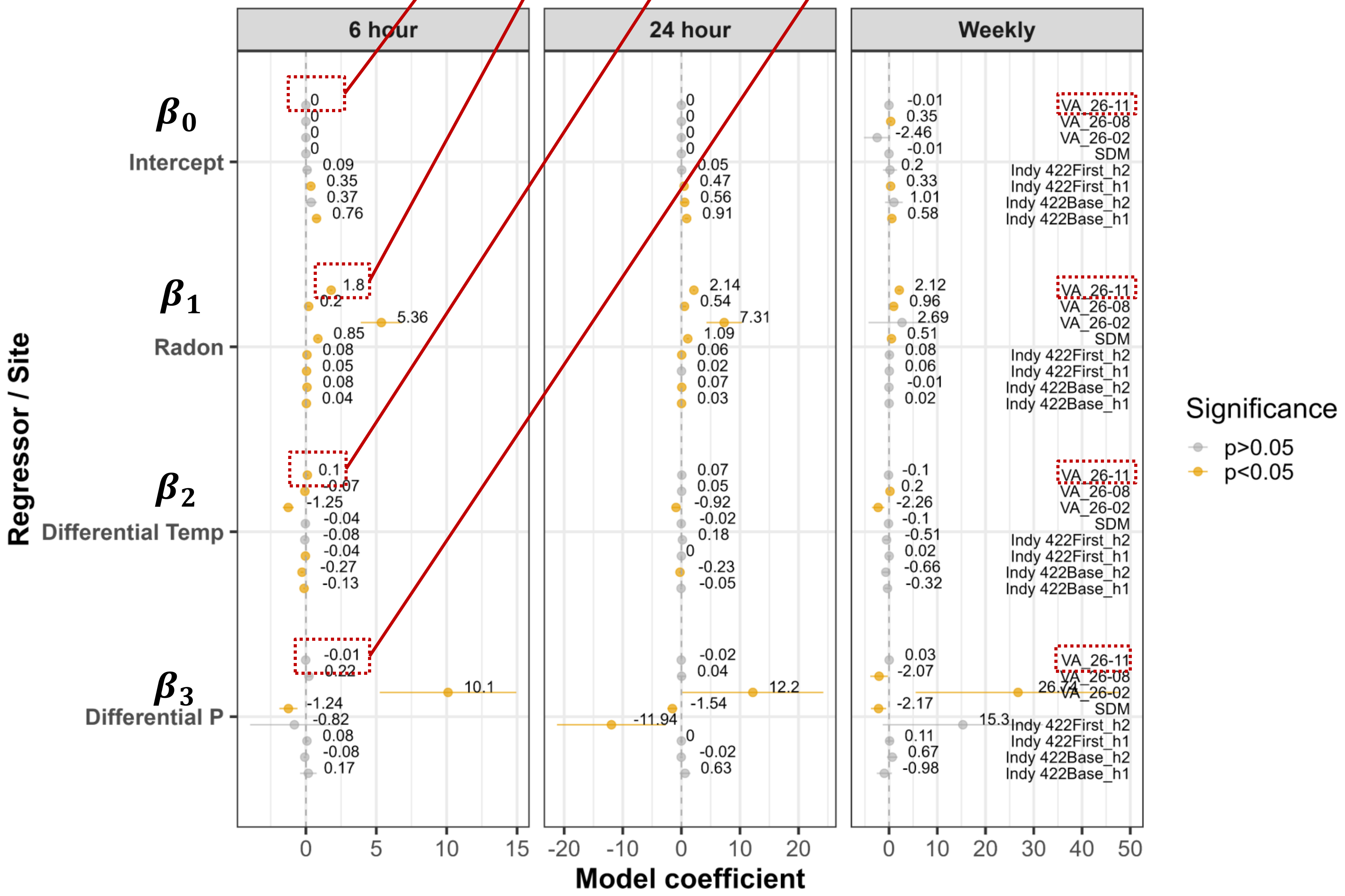
Putting it all together: Coefficients for all 24 **complete multivariate models**

$$C_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} R_t + \boldsymbol{\beta_2} \Delta T_t + \boldsymbol{\beta_3} \Delta P_t + \eta_t$$

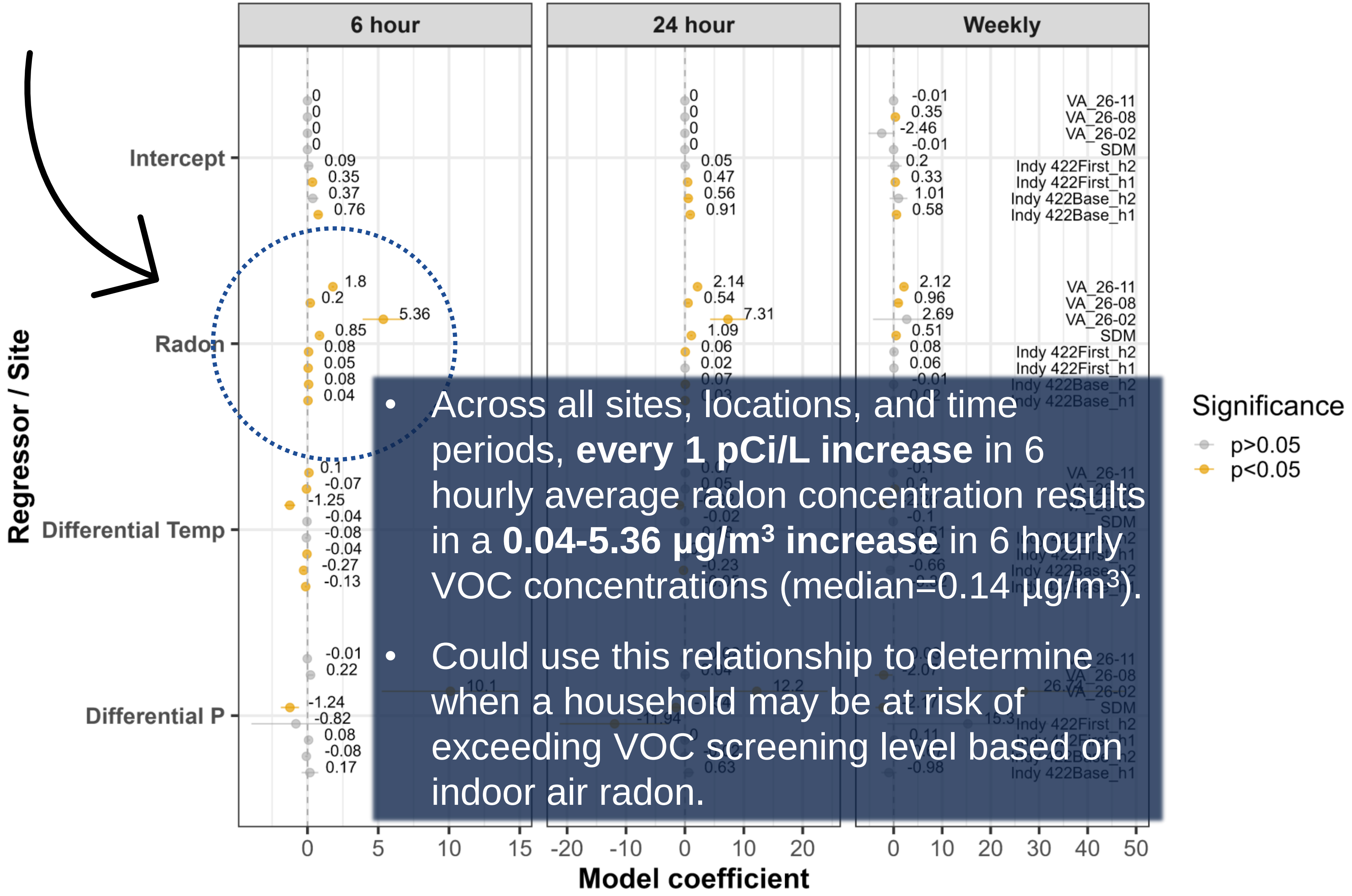


Putting it all together: Coefficients for all 24 **complete multivariate models**

$$C_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} R_t + \boldsymbol{\beta_2} \Delta T_t + \boldsymbol{\beta_3} \Delta P_t + \eta_t$$

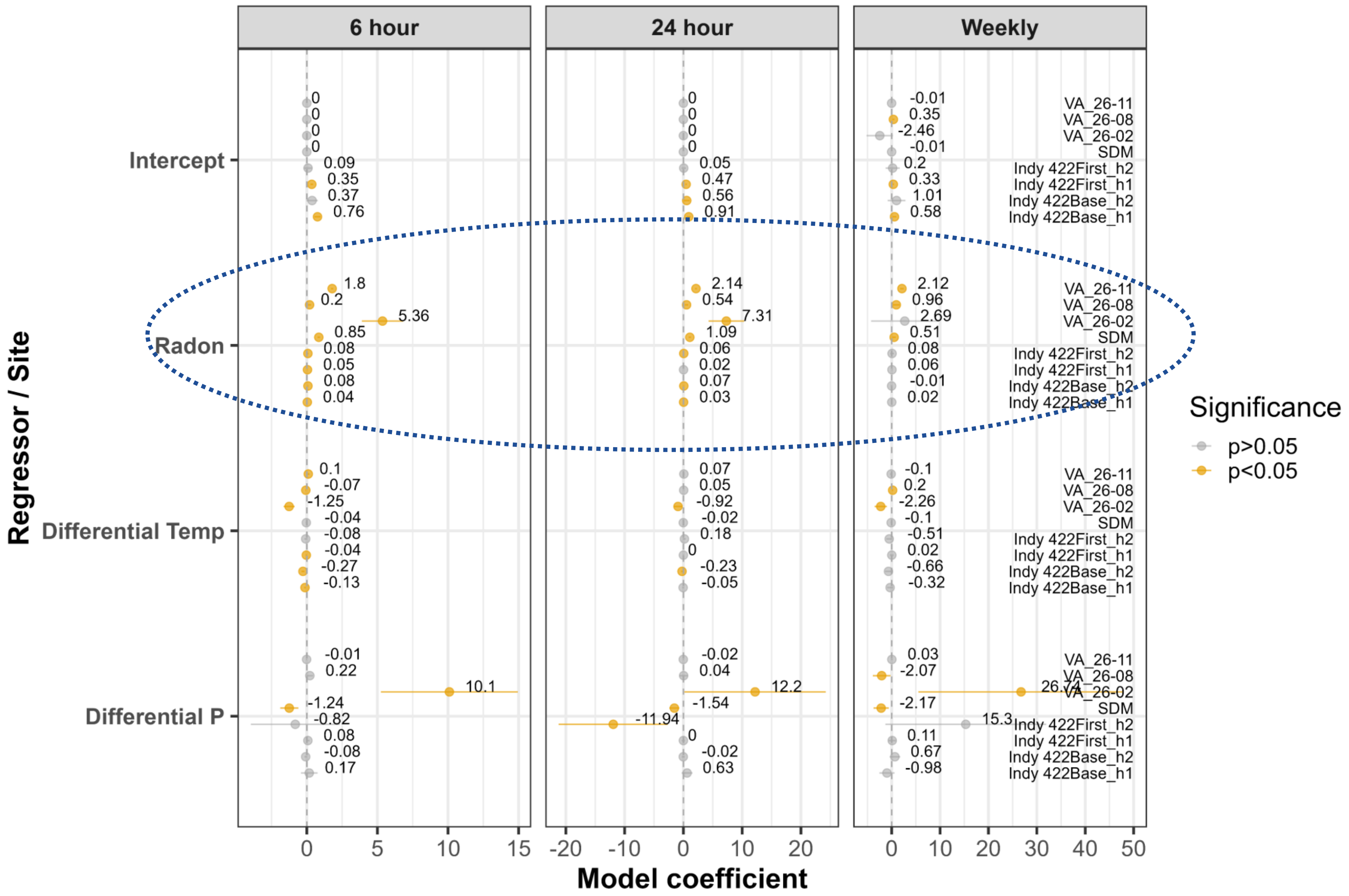


Six hourly radon concentrations significant across **all sites**, sample **locations**, and **time periods**

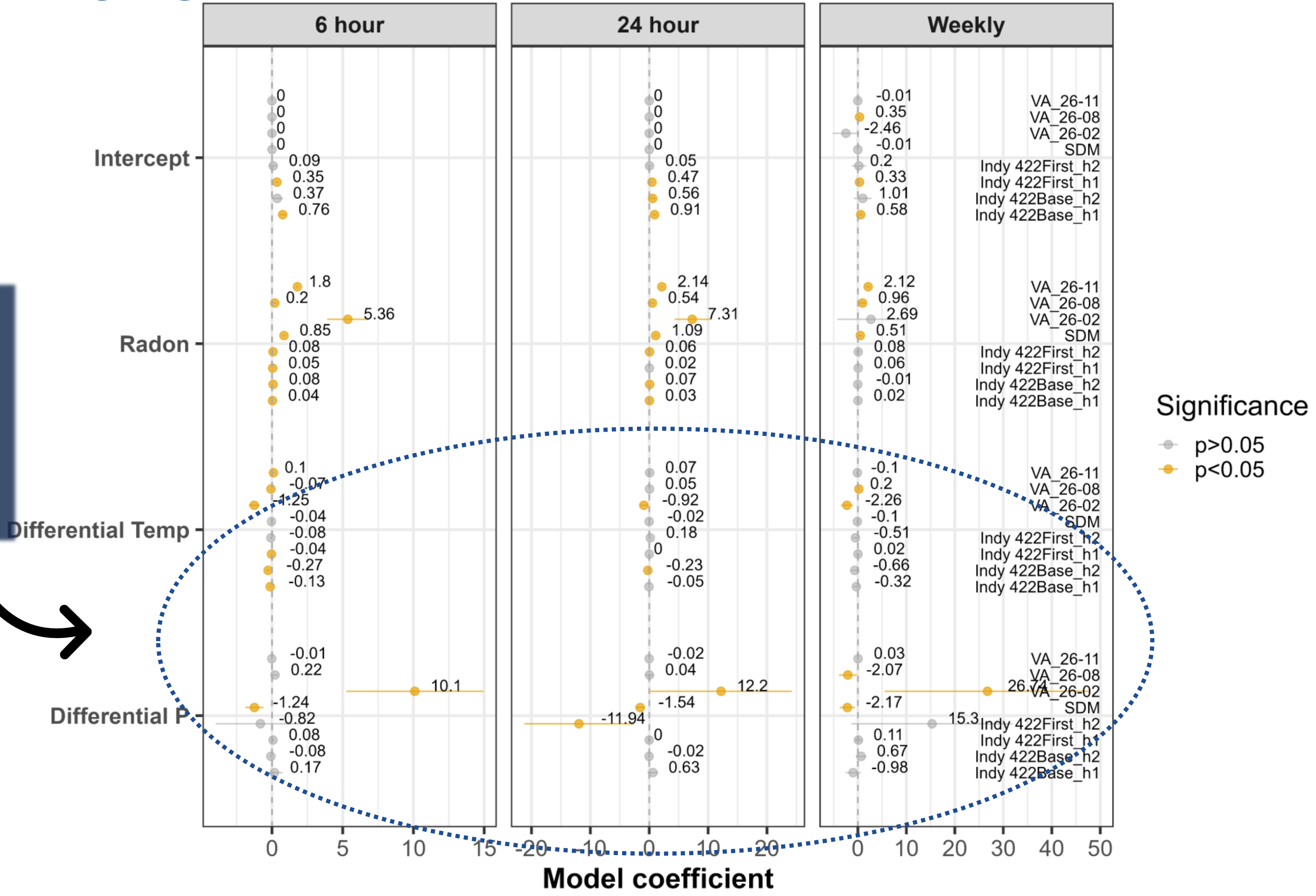


- Across all sites, locations, and time periods, **every 1 pCi/L increase** in 6 hourly average radon concentration results in a **0.04-5.36 µg/m³ increase** in 6 hourly VOC concentrations (median=0.14 µg/m³).
- Could use this relationship to determine when a household may be at risk of exceeding VOC screening level based on indoor air radon.

Radon **less reliable** as a tracer at **longer averaging times**



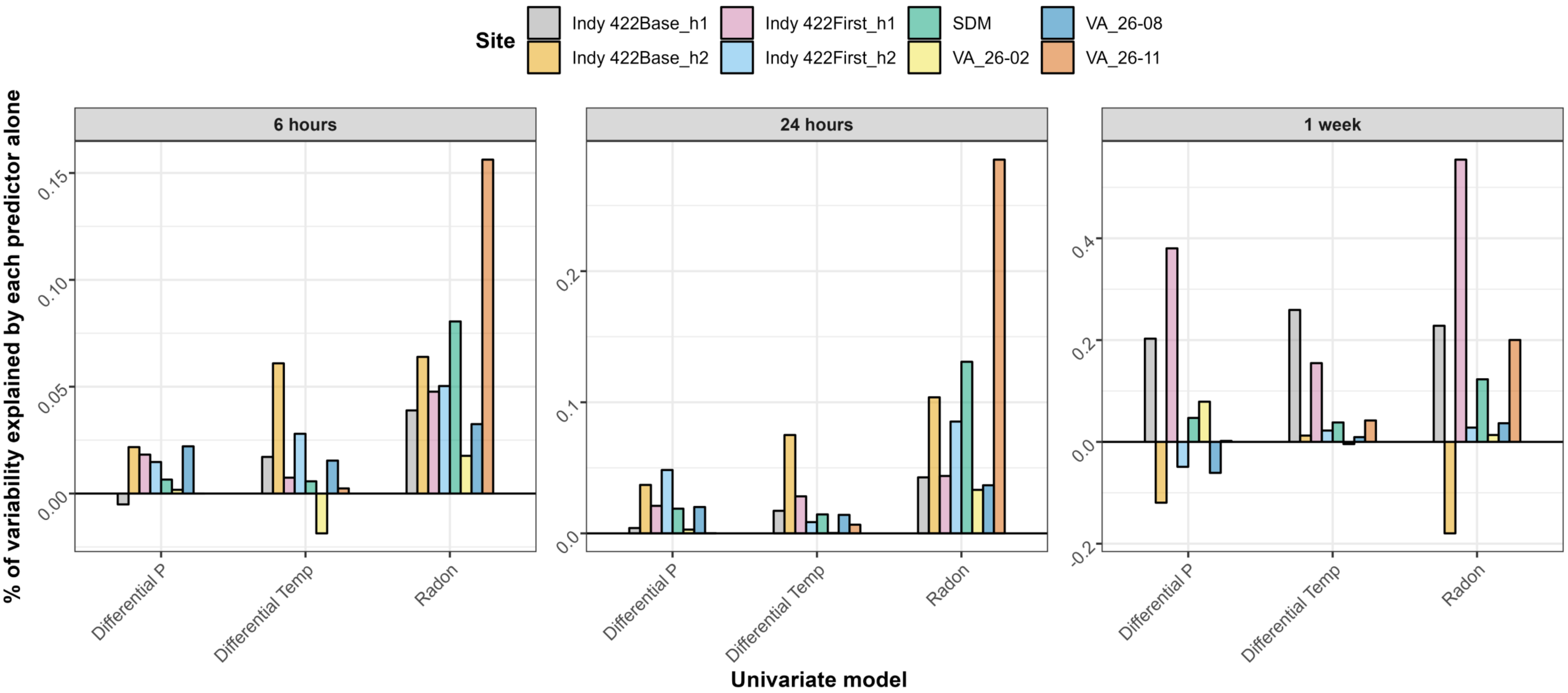
Significance of other indicators and tracers **site specific** and **dependent on sample averaging time**



Differential pressure and temperature dependent on the **sample location** within the building and the **sampling period**.

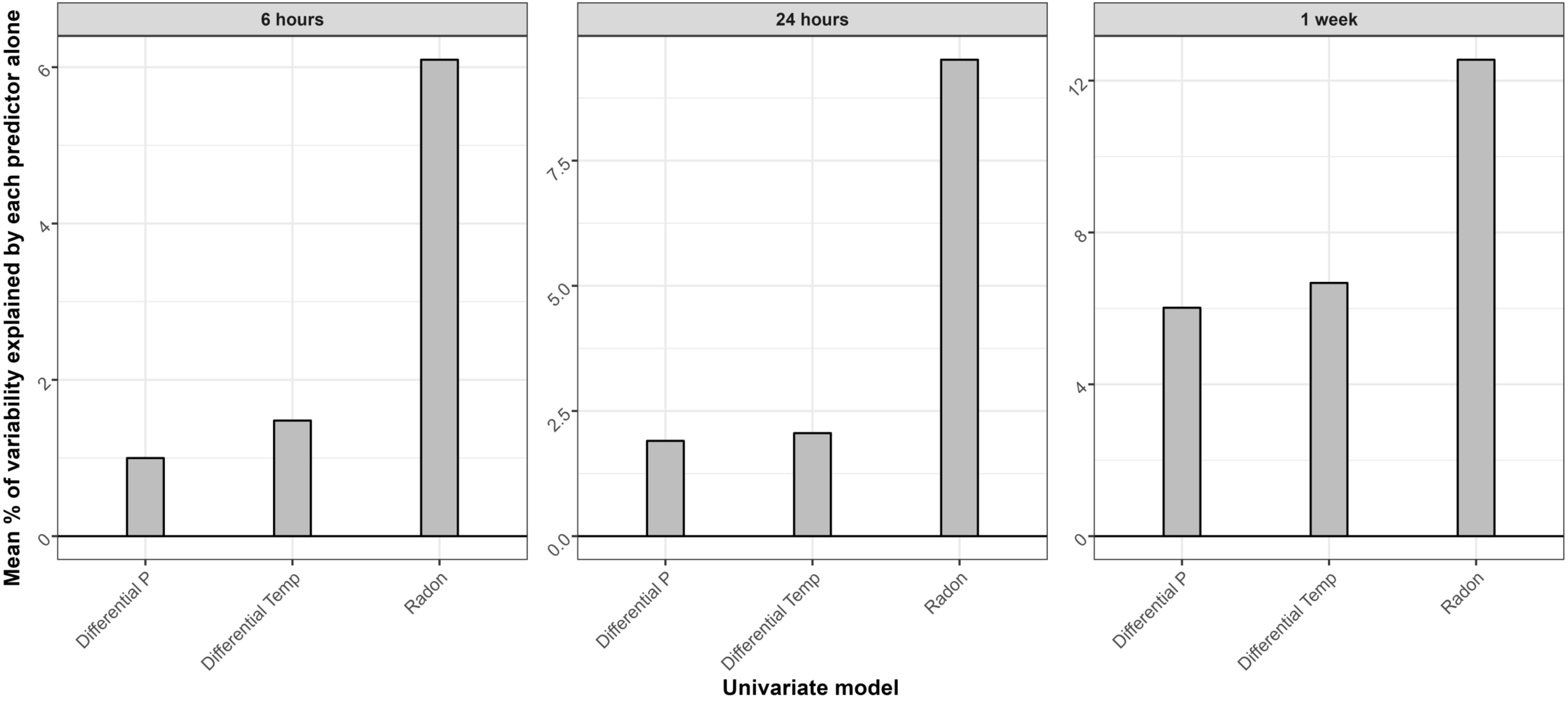
Variable importance – How much does the fit improve with a univariate model compared to a *null model*?

Greater increase = greater importance



Variable importance – How much does the fit improve with a univariate model compared to a *null model*?

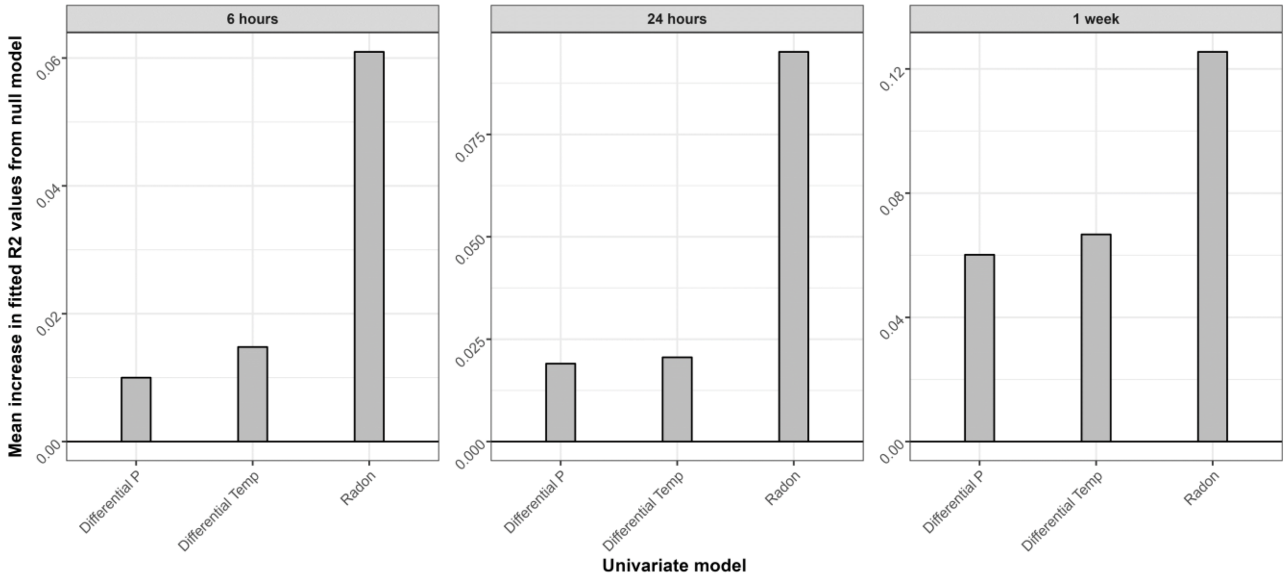
Greater increase = greater importance



Variable importance – How much does the fit improve with a univariate model compared to a *null model*?

Greater increase in R^2 = greater importance

- On average, **radon independently explains an additional 6-12% of the variability** in VOC concentrations beyond the temporal autocorrelation alone.
- **Differential temperature and differential pressure each explain only 1-7% of the variability** in VOC concentrations beyond the temporal autocorrelation alone.



Conclusions & Next steps

- **6-hour radon concentration is the most reliable tracer** for indoor air VOC concentrations across sites after controlling for temporal autocorrelation
  - Relationship between Δ6-hour radon and Δ6-hour VOC varies by site
  - Baseline relationship could be characterized to develop general recommendations for different areas
  - Simple in-home radon detectors may be the best, most immediate indicator of increased VOC exposure risk in VI areas

- **Other indicators are largely site specific** and vary according to location within the building and time of year

- **Future analysis:** Include additional covariates, include additional sites, forecasting future VI

Thank you

Contact: Riley Mulhern | email: rmulhern@rti.org