# spatialsample:

## A tidy approach to spatial cross-validation

**Michael J Mahoney**

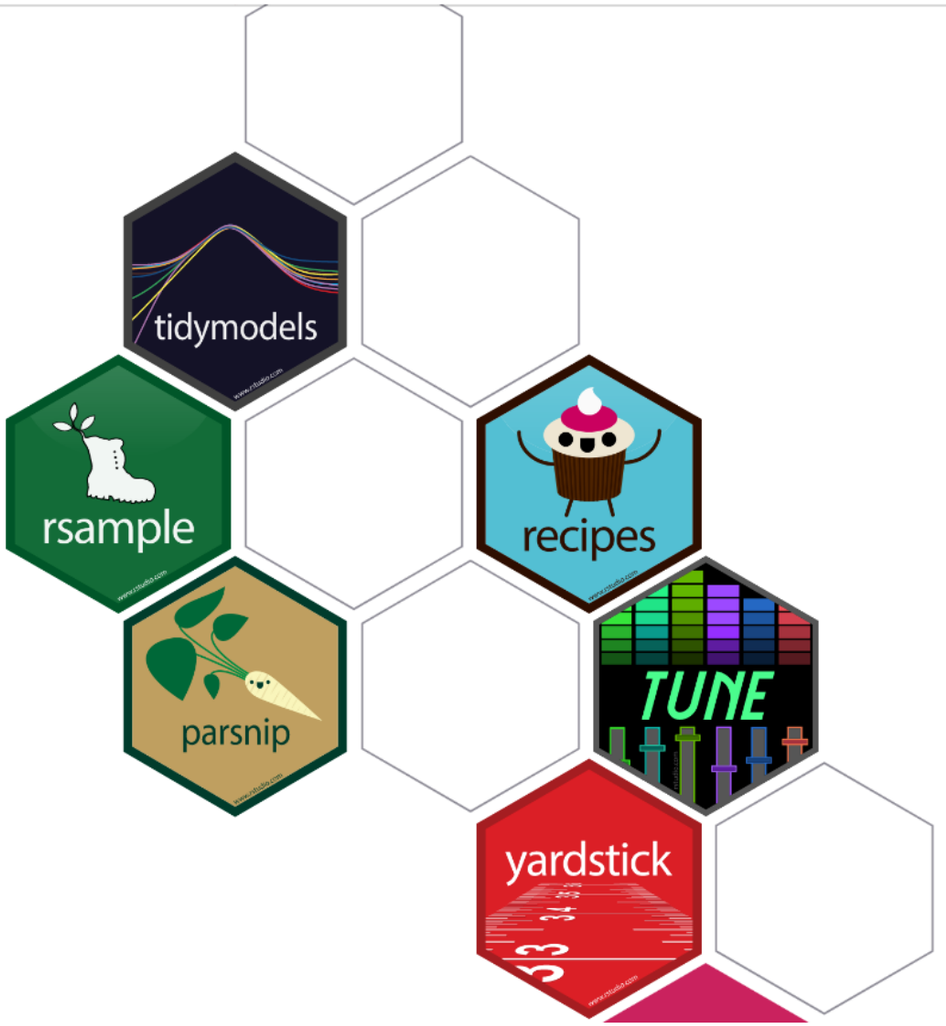
mjmahone@esf.edu

# About Me

- Mike Mahoney
- PhD candidate in environmental science
- 2022 summer intern with Posit (spatialsample, rsample)
- These slides: mm218.dev/boston_useR_2023

Tidymodels

PACKAGES GET STARTED LEARN HELP CONTRIBUTE



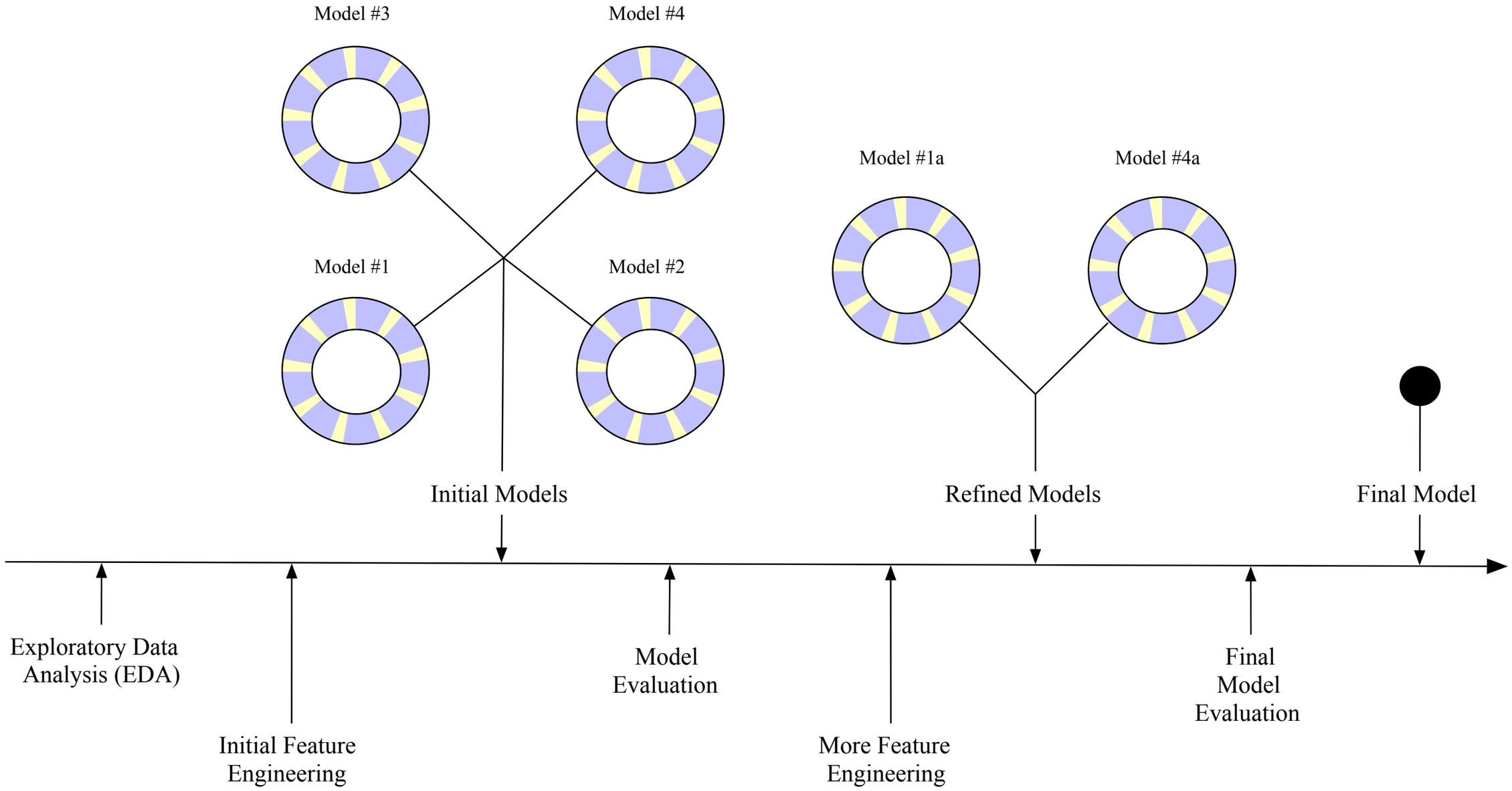

## TIDYMODELS

The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.

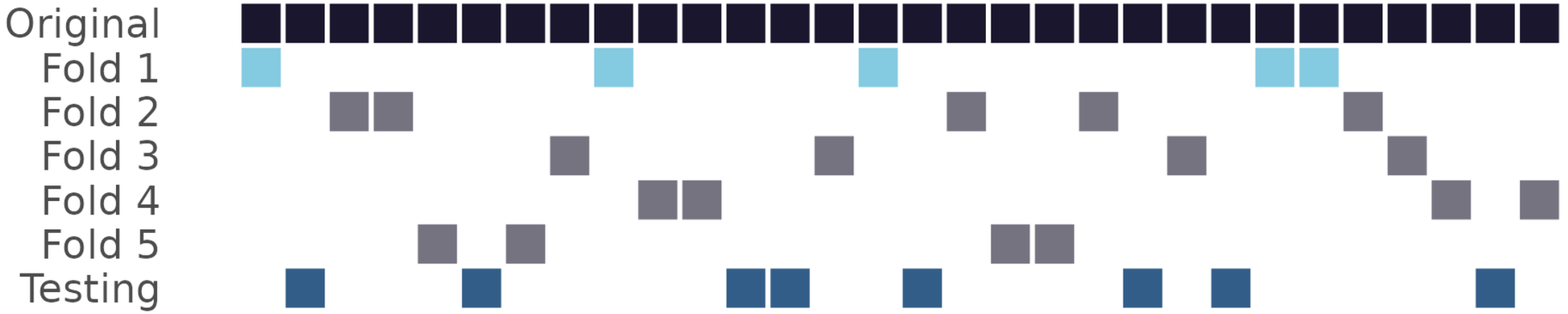
Install tidymodels with:

```
install.packages("tidymodels")
```

# Data splitting:


Model #3
Model #4
Model #1
Model #2
Model #1a
Model #4a
Initial Models
Refined Models
Final Model
Exploratory Data Analysis (EDA)
Initial Feature Engineering
Model Evaluation
More Feature Engineering
Final Model Evaluation

# Cross-validation:



# rsample and friends

```r
library(tidymodels)
rsample::vfold_cv(spatialsample::boston_canopy) |> head()
```

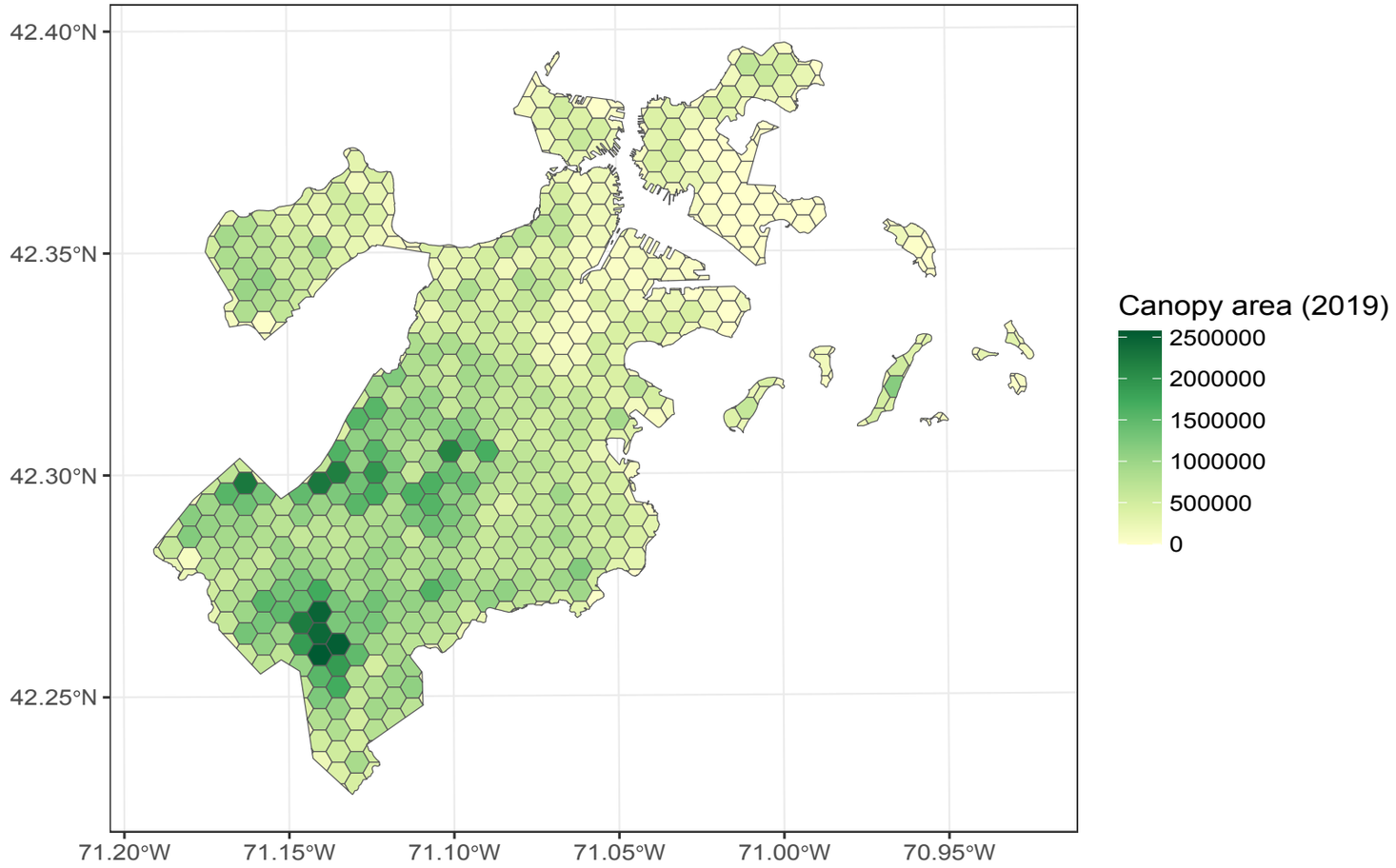
```
#> # A tibble: 6 × 2
#>   splits           id
#>   <list>           <chr>
#> 1 <split [613/69]> Fold01
#> 2 <split [613/69]> Fold02
#> 3 <split [614/68]> Fold03
#> 4 <split [614/68]> Fold04
#> 5 <split [614/68]> Fold05
#> 6 <split [614/68]> Fold06
```

```r
workflow() |>
  add_model(linear_reg()) |>
  add_formula(canopy_area_2019 ~ land_area * mean_temp) |>
  fit_resamples(vfold_cv(spatialsample::boston_canopy)) |>
  collect_metrics()
```

```
#> # A tibble: 2 × 6
#>   .metric .estimator       mean     n    std_err .config
#>   <chr>   <chr>           <dbl> <int>      <dbl> <chr>
#> 1 rmse    standard   377089.       10 20426.     Preprocessor1_Model1
#> 2 rsq     standard        0.353    10     0.0178 Preprocessor1_Model1
```

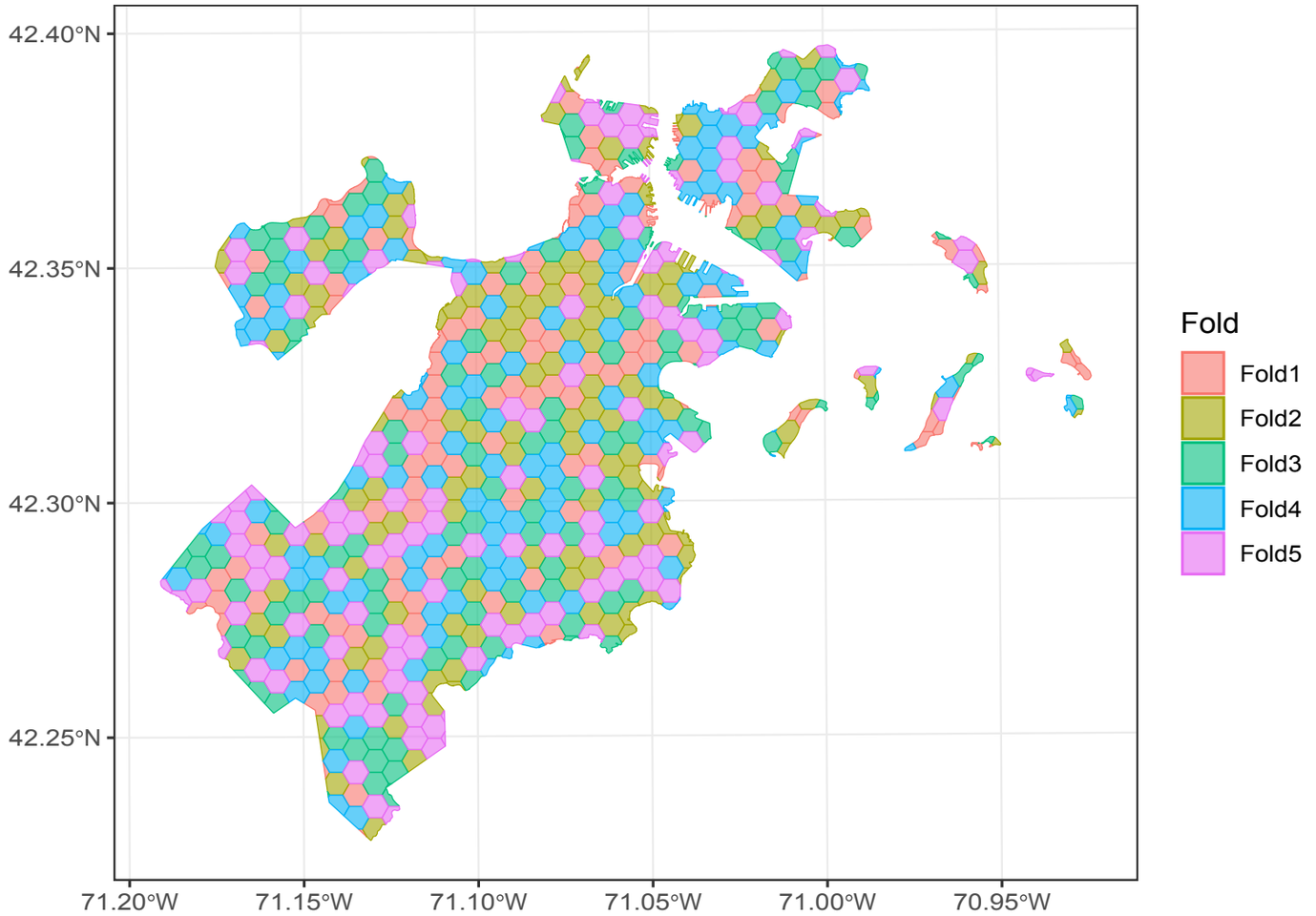
# What does “new data” mean?

```
1 ggplot(spatialsample::boston_canopy, aes(fill = canopy_area_2019)) + geom_sf() +
2   scale_fill_distiller(name = "Canopy area (2019)", palette = "YlGn", direction = 1)
```



# Are these folds really unrelated?

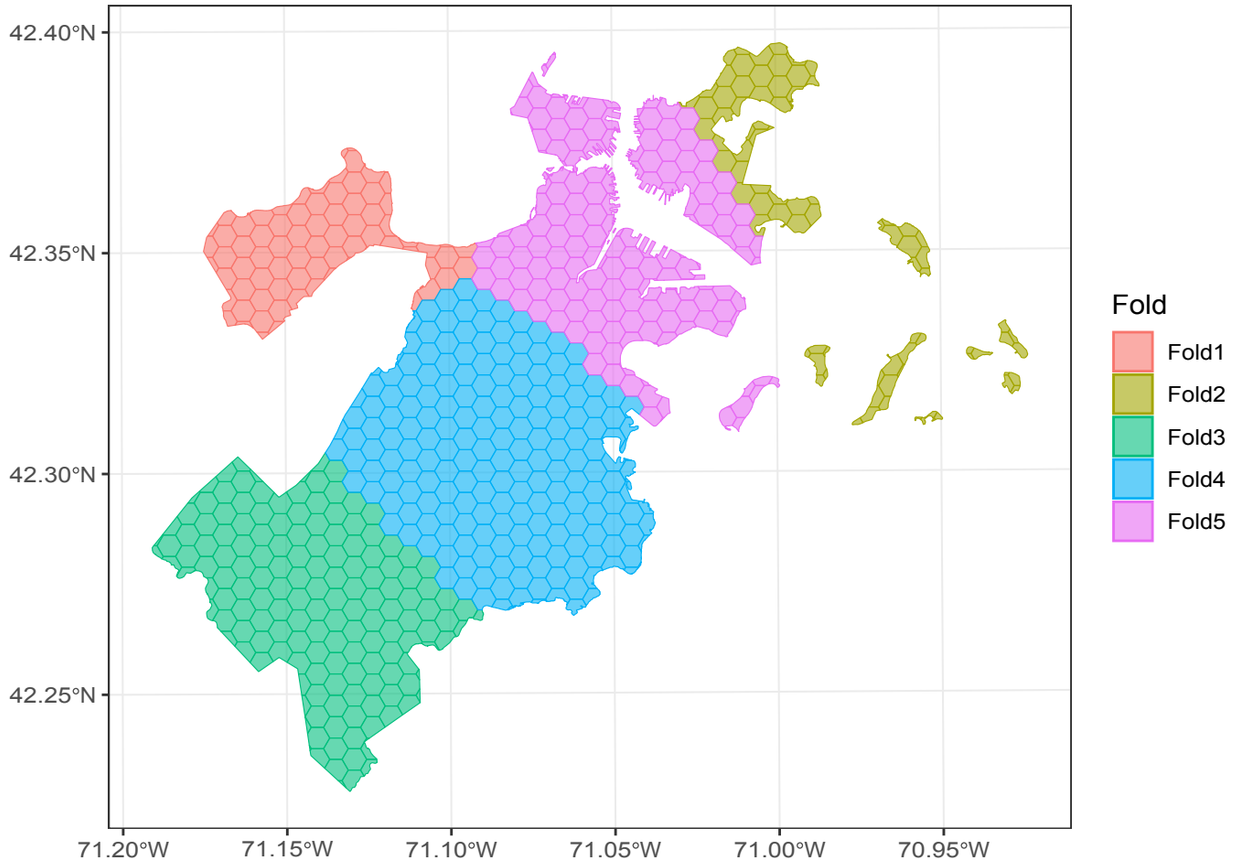
```
1  rsample::vfold_cv(spatialsample::boston_canopy, v = 5)
```




spatialsample

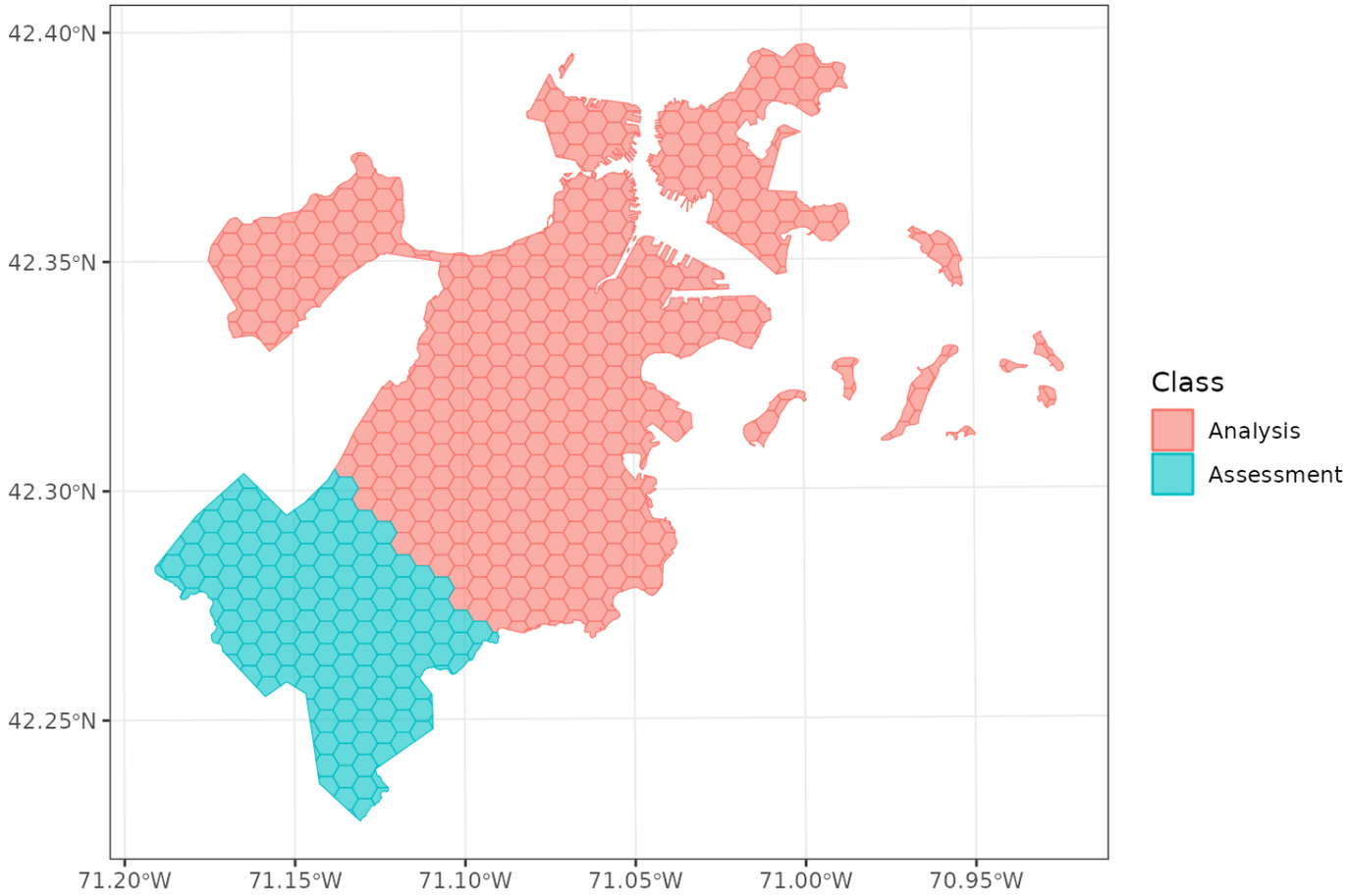
# Spatial clustering

```
1 library(spatialsample)
2 set.seed(1234)
3 spatial_clustering_cv(boston_canopy, v = 5)
```



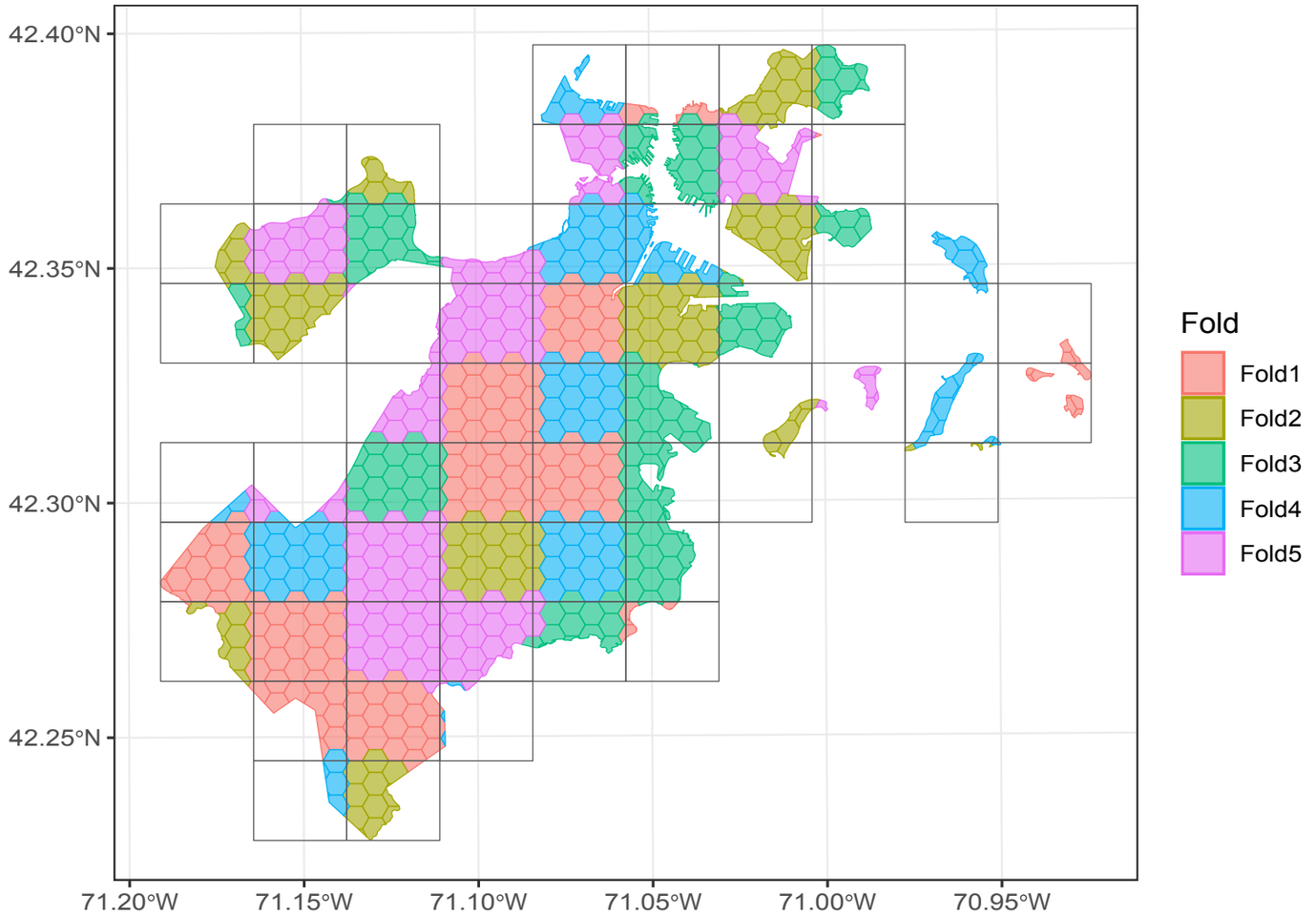
# Spatial clustering

```
1 library(purrr)
2 walk(spatial_clustering_cv(boston_canopy, v = 5)$splits, function(x) print(autoplot(x)))
```



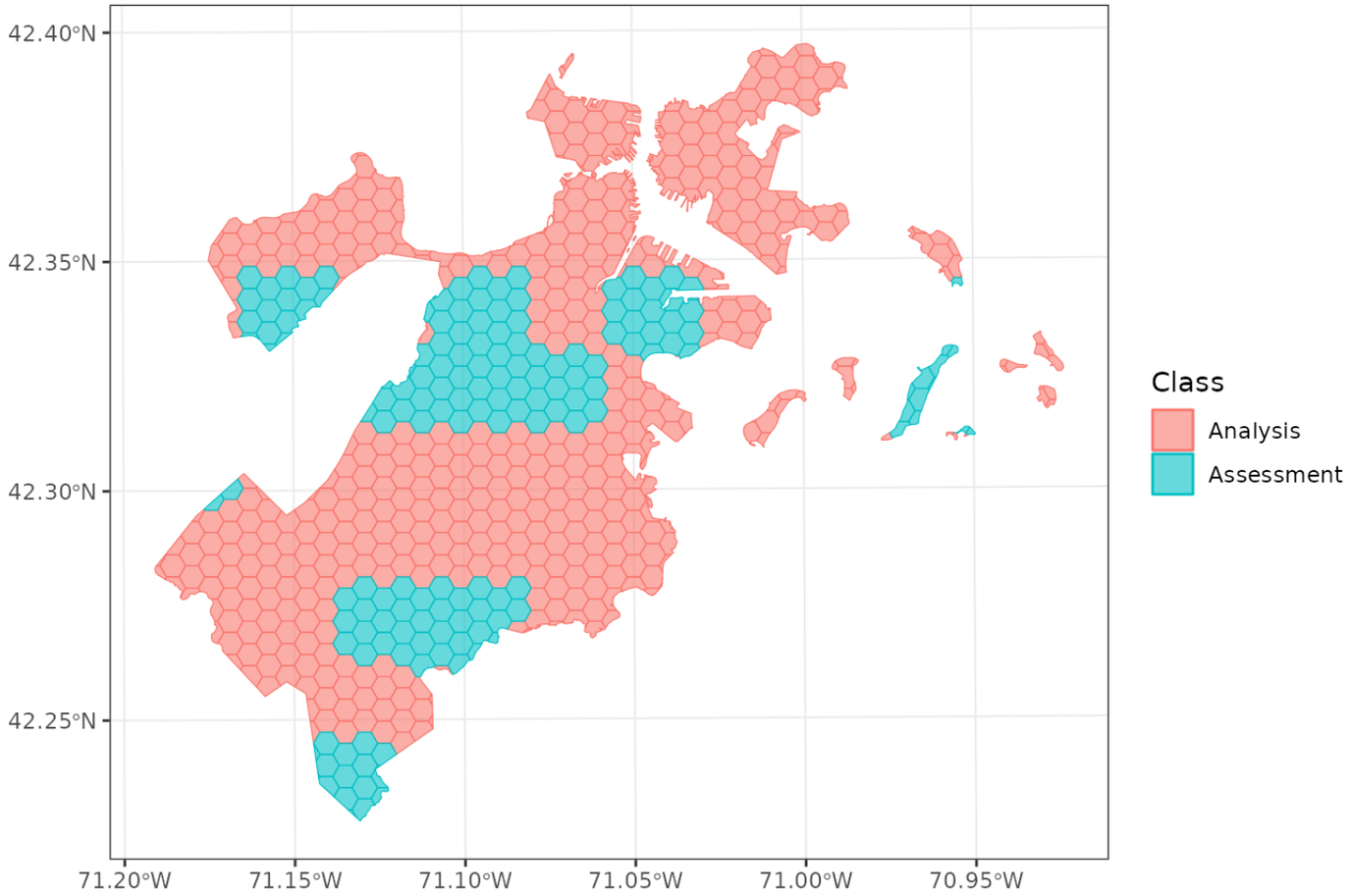
# Spatial blocking

```
1  spatial_block_cv(boston_canopy, v = 5, n = c(10, 10))
```



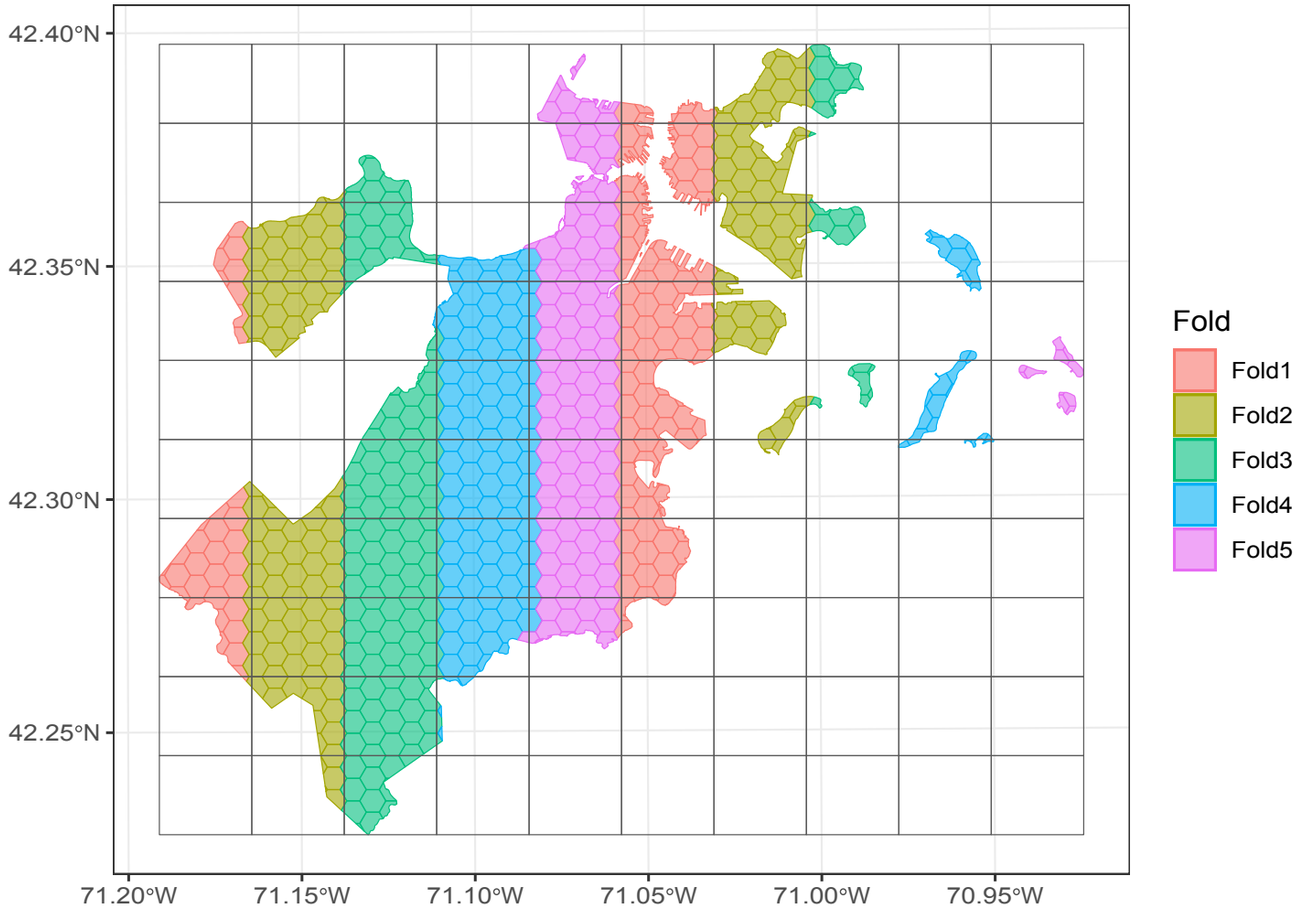
# Spatial blocking

```
walk(spatial_block_cv(boston_canopy, v = 5, n = c(10, 10))$splits,
     function(x) print(autoplot(x)))
```



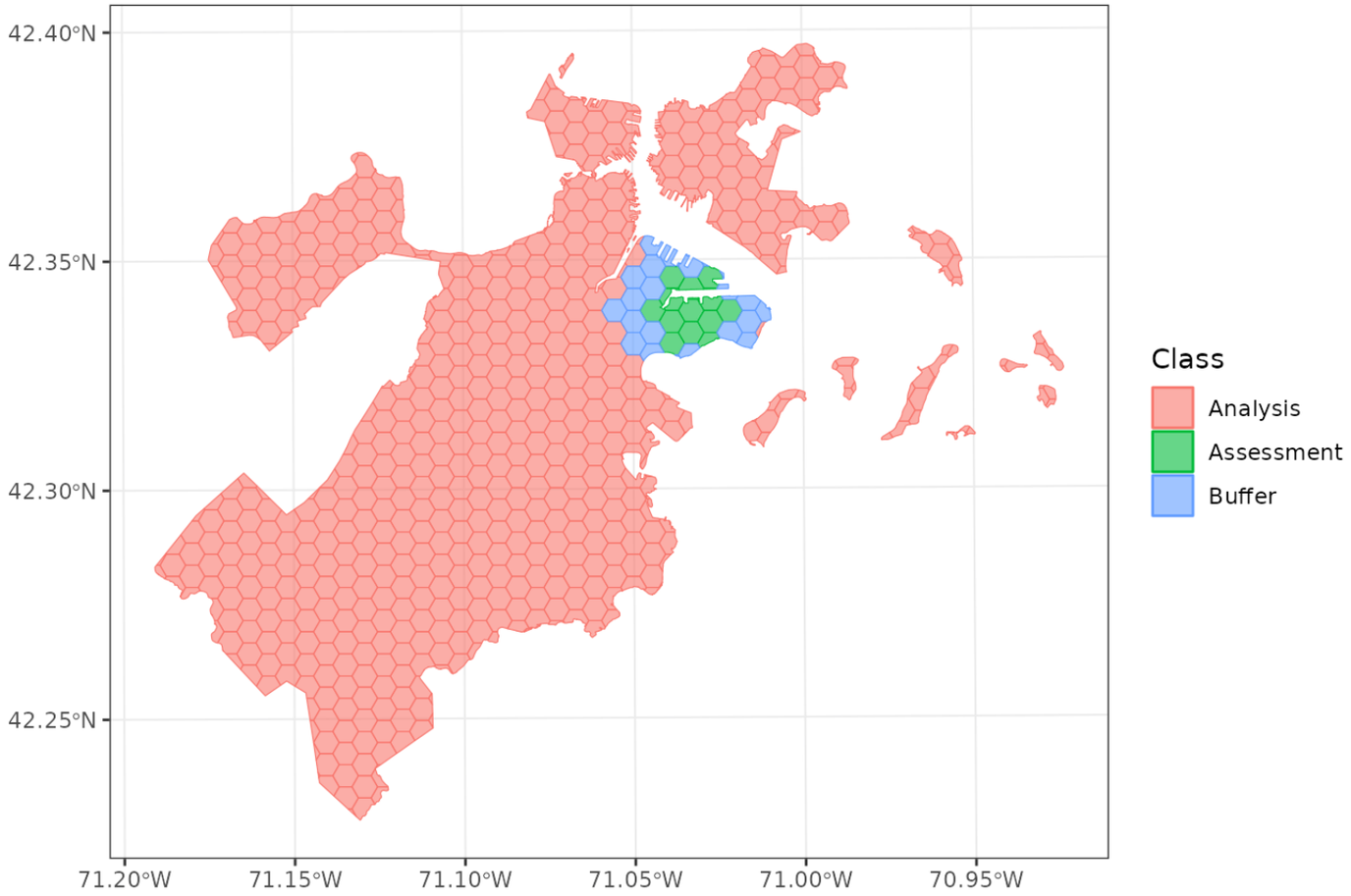
# Spatial blocking

```
spatial_block_cv(boston_canopy, v = 5, n = c(10, 10),
                 method = "continuous", relevant_only = FALSE)
```



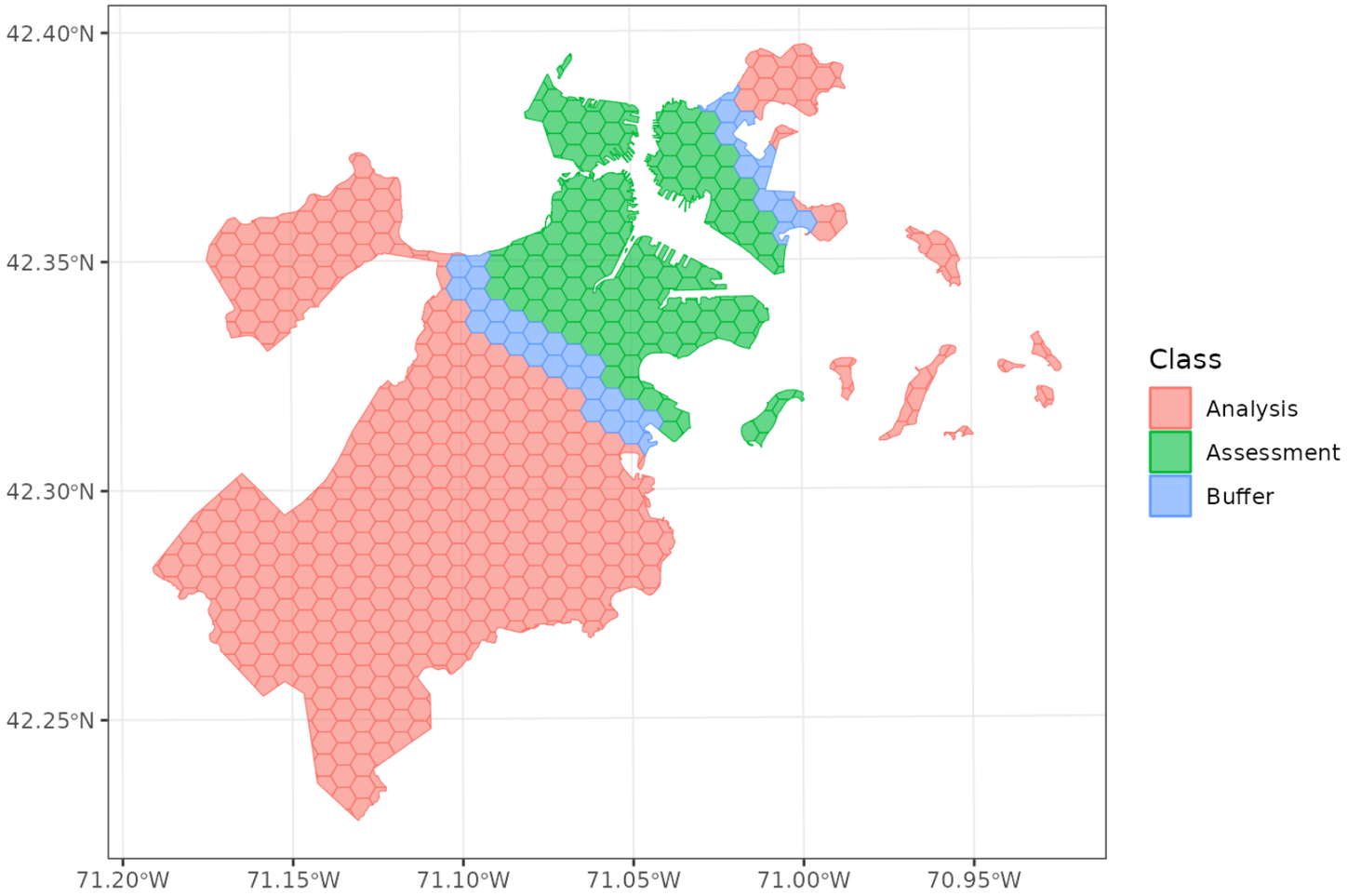
# Spatial LODO

```
folds <- spatial_buffer_vfold_cv(boston_canopy, v = Inf, radius = 1500, buffer = 1500)
walk(folds$splits, function(x) print(autoplot(x)))
```



# Buffering

```
1  spatial_clustering_cv(boston_canopy, v = 5, buffer = 1500)
```



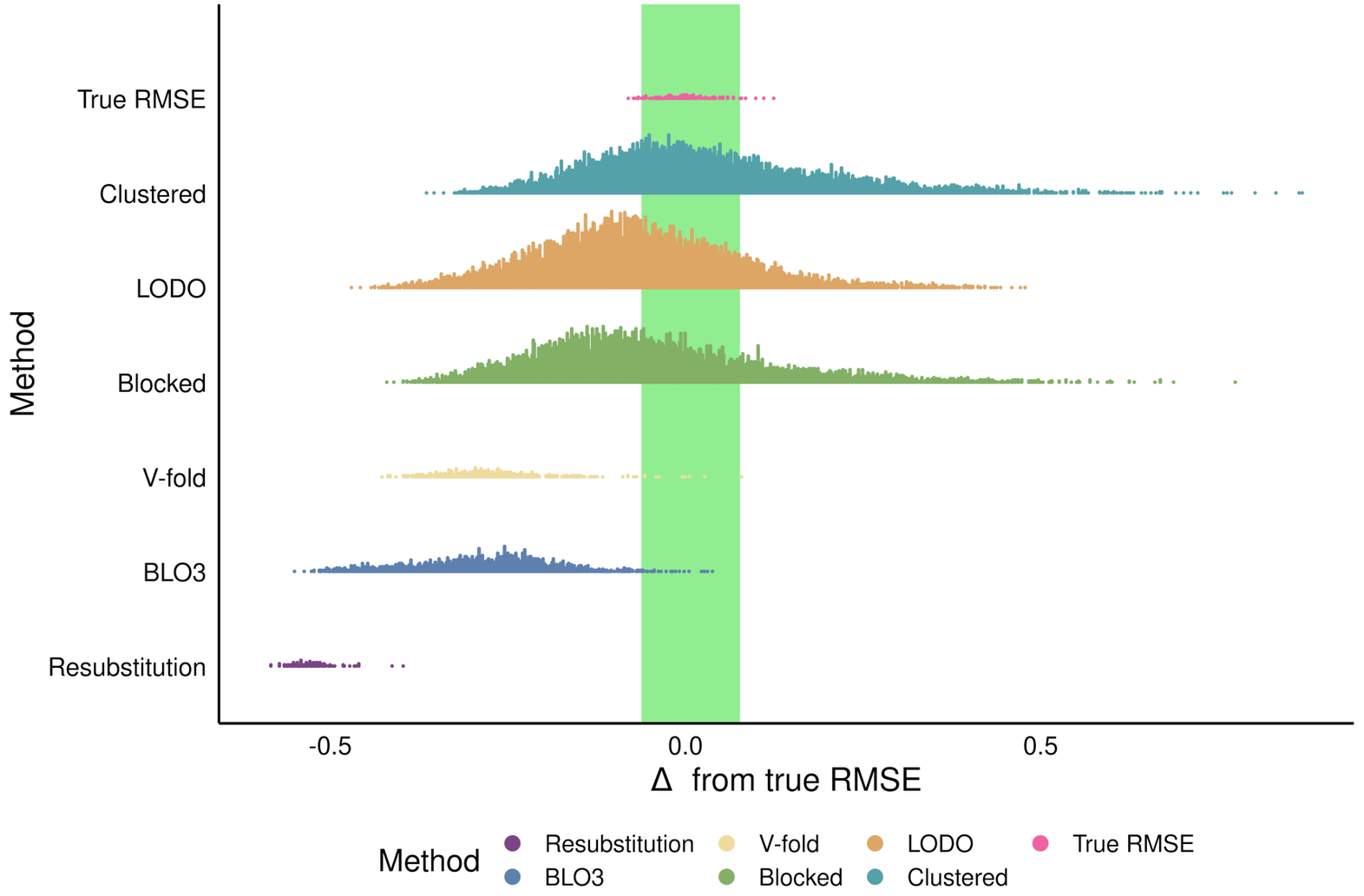
# tidymodels integration

```r
workflow() |>
  add_model(linear_reg()) |>
  add_formula(canopy_area_2019 ~ land_area * mean_temp) |>
  fit_resamples(vfold_cv(spatialsample::boston_canopy)) |>
  collect_metrics()
```

```
#> # A tibble: 2 × 6
#>   .metric .estimator       mean     n    std_err .config
#>   <chr>   <chr>           <dbl> <int>      <dbl> <chr>
#> 1 rmse    standard   378993.       10 18001.     Preprocessor1_Model1
#> 2 rsq     standard        0.354    10     0.0171 Preprocessor1_Model1
```

```r
workflow() |>
  add_model(linear_reg()) |>
  add_formula(canopy_area_2019 ~ land_area * mean_temp) |>
  fit_resamples(spatial_clustering_cv(spatialsample::boston_canopy)) |>
  collect_metrics()
```

```
#> # A tibble: 2 × 6
#>   .metric .estimator       mean     n     std_err .config
#>   <chr>   <chr>           <dbl> <int>       <dbl> <chr>
#> 1 rmse    standard   934397.       10 549480.     Preprocessor1_Model1
#> 2 rsq     standard        0.348    10      0.0468 Preprocessor1_Model1
```




Mahoney, MJ, Johnson, L. K., Silge, J., Frick, H., Kuhn, M., and Beier, C. M. In Review. Assessing the performance of spatial cross-validation approaches for models of spatially structured data. https://doi.org/10.48550/arXiv.2303.07334

# Other features:

- ☑ Works with projected & geographic CRS
- ☑ Arguments accept explicit units
- ☑ Aware of CRS units, functions do unit conversion
- ☑ Handles all geometry types*

# Thank you!

**Find me online:**

mm218.dev

@mikemahoney218

@MikeMahoney218@fosstodon.org

Slides available at mm218.dev/boston_useR_2023

More spatialsample: https://spatialsample.tidymodels.org/