Project Draft Models

John Hong, Tomoki Takeuchi, Jose Hernandez, David Martinez Vasquez

10/30/2021

Dataset

```
birds <- read.csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQwR56LrrsuT8BMSLbN6bmu8CPLIFhX5LauA

birds <- birds %>% mutate(
  Year = factor(Year),
  Property = factor(Property),
  UTM_E = UTM_E / 1000000,
  UTM_N = UTM_N / 1000000
)
```

```
#birds
```

Model

Model Plan

Response variable GCW because we are interested in the bird observation.

Predictor variables We are going to use the following predictors.

- **MeanImp**: Percentage that impervious surface accounts for within each cell
- **Year**: Year when a bird is observed
- **Property**: Name of a preserve in which birds are observed
- **UTM_E**: x Coordinate in meters divided by 1,000,000
- **UTM_N**: y Coordinate in meters divided by 1,000,000

(UTM_E and UTM_N are divided by 1,000,000 to reduce the scale)

- `MeanImp` because we are interested in how development would affect the bird habitation.
- `Year` because we want to see the transition over time.
- `Property` because we want to see how different preserves associate with bird observation.
- `UTM_E` and `UTM_N` because we are interested in the specific locations of bird observation.

Model Fitting

First, we used negative binomial model to fit the data, but the ACF looked really bad. Next, we tries using zero-inflation model, but again, it failed at the ACF condition check. We also tried using Generalized Additive Models, but it did not work as well because of the auto-correlation.

```r
# One of the models we tried but failed
#bird_nb <-
#      glmmTMB(GCW ~ MeanImp + (1|Year) + (1|Property) +
#                UTM_N + UTM_E,
#      family = nbinom1(link = 'log'),
#      na.action = 'na.fail',
#      data = birds)
```

After those tries, we decided to use Generalized Additive Mixed Models to build a 2D-Smooth model. The reasons we decided to use gamm are:

- there are multiple hot spots of bird observations at different x coordinate and y coordinate.
- we want to smooth variables.
- we want to add random effect to `Year` and `Property`.
- we want to add spatial autocorrelation in the residuals.

```r
bird_gam <- gamm(GCW ~ s(MeanImp, k = 10, bs = 'cs') +
                      te(UTM_E, UTM_N, bs = 'ts', k = 11),
                      # s(UTM_E, UTM_N, bs = 'gp', k = 100, m = 2),
                    random = list(Year = ~1, Property = ~1),
                    correlation = corSpatial(form = ~ UTM_E + UTM_N, type='gaussian'),
                    data = birds,
                    method = "ML",
                    family = negbin(theta = 0.6, link = 'log'))
```

```
##
##  Maximum number of PQL iterations:  20

## iteration 1

## iteration 2

## iteration 3

## iteration 4

## iteration 5

## iteration 6

## iteration 7

## iteration 8

## iteration 9

## iteration 10

## iteration 11
```

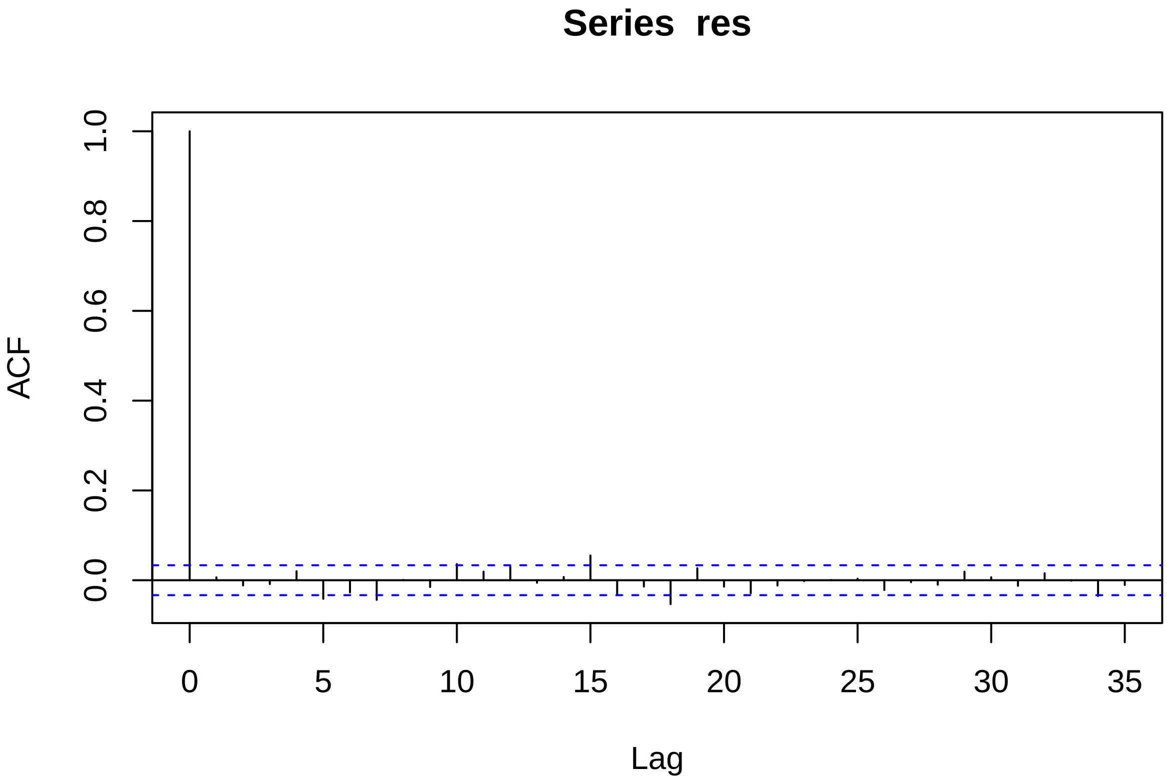
```r
# calculate residuals
res <- resid(bird_gam$lme, type = 'normalized') %>%
  as.numeric()

fitted <- fitted(bird_gam$gam, type = 'normalized') %>%
  as.numeric()
```

Model Assesment

It acceptably meets model conditions.

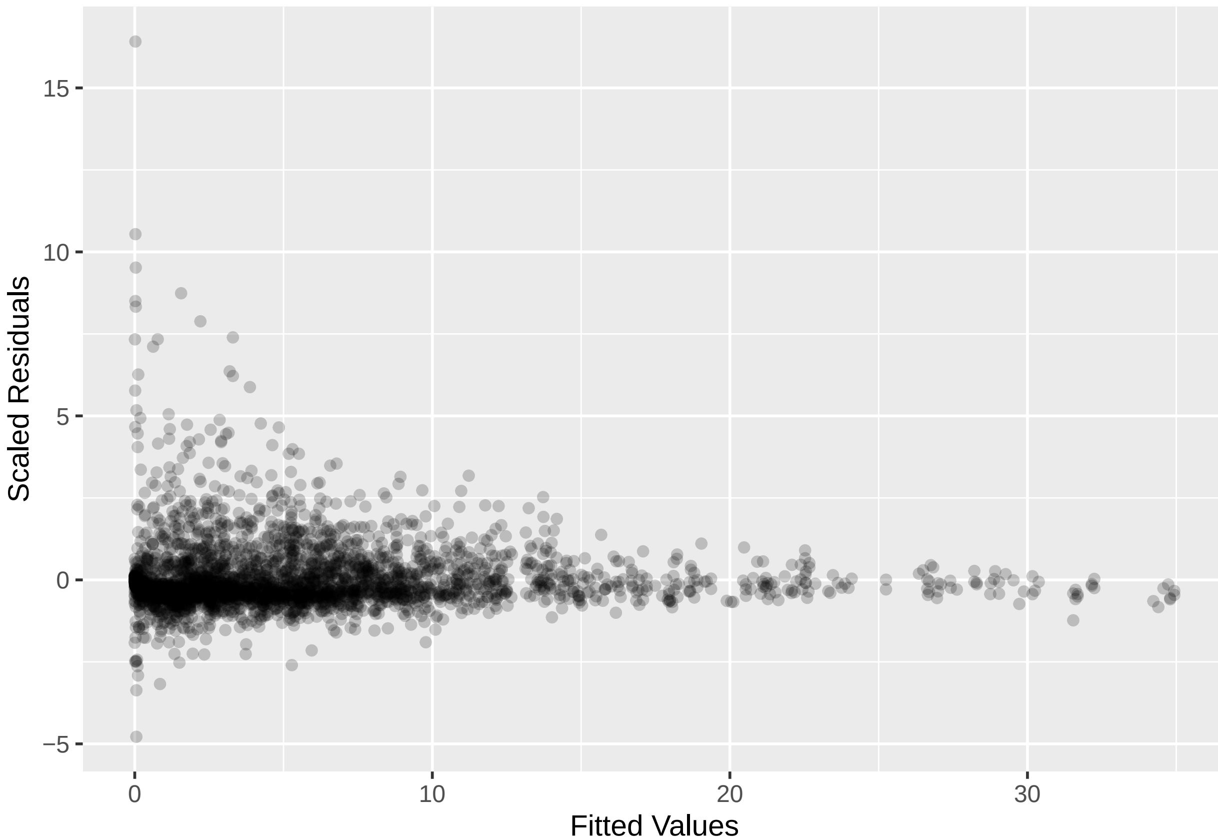
Auto-Correlation of Residuals This graph checks the auto-correlation of residuals. It passes the test since they are not protruding much out of the confidence bounds.

```
acf(res)
```



Scaled Residuals Condition This graph checks mean-variance condition. According to this plot, although the number of points decreases as fitted value gets larger, there is no evidence of non-constant variance. It does not look like there is a linear or nonlinear trend, and thus, linearity condition is satisfied as well.

```
gf_point(res ~ fitted, alpha = 0.2) %>%
  gf_labs(x = 'Fitted Values', y = 'Scaled Residuals')
```



Graphics

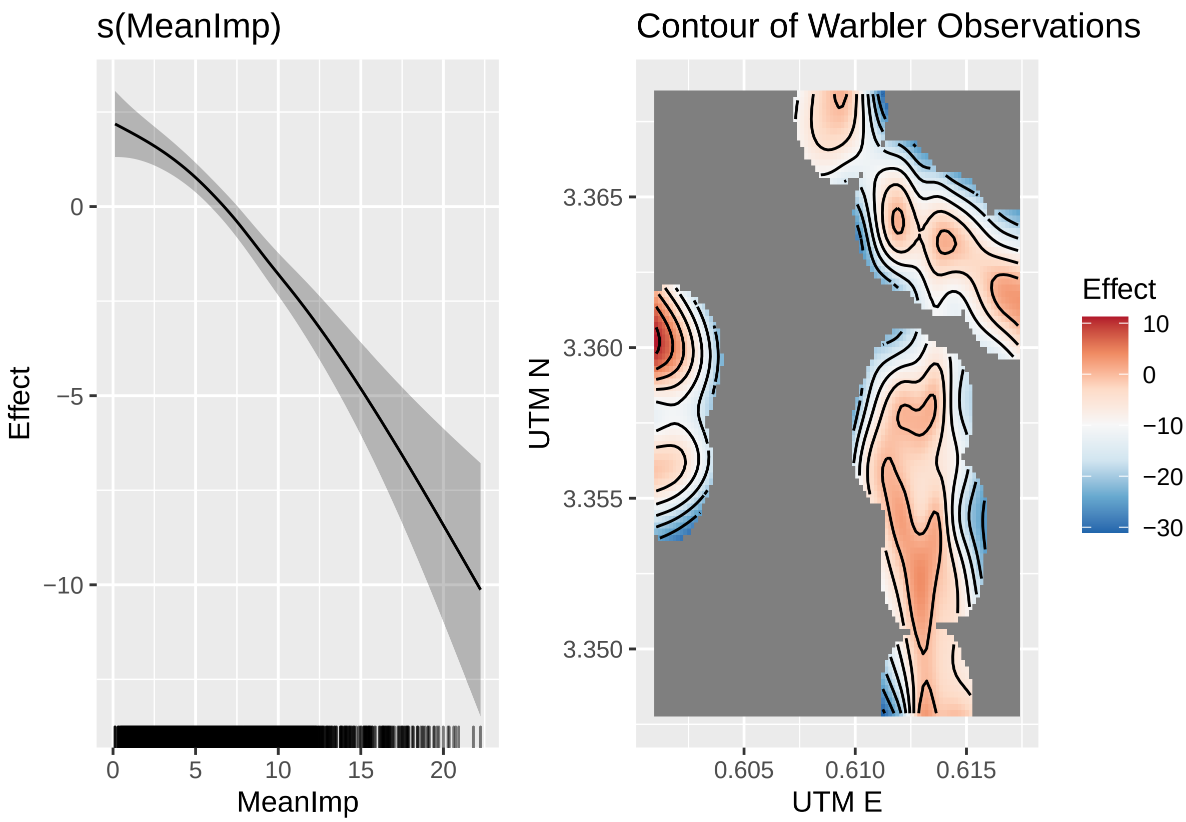
This graph shows that the effect of one model term.

As MeanImp increases, the number of bird observations decreases almost linearly, and when the MeanImp is about 20%, the least number of birds was observed. There is a difference of about 12 birds between areas with 0% and 20% impervious surface.

The map shows the density distribution of birds. Red areas are where there are more birds than average. In contrast, blue areas are where there are less birds than average.

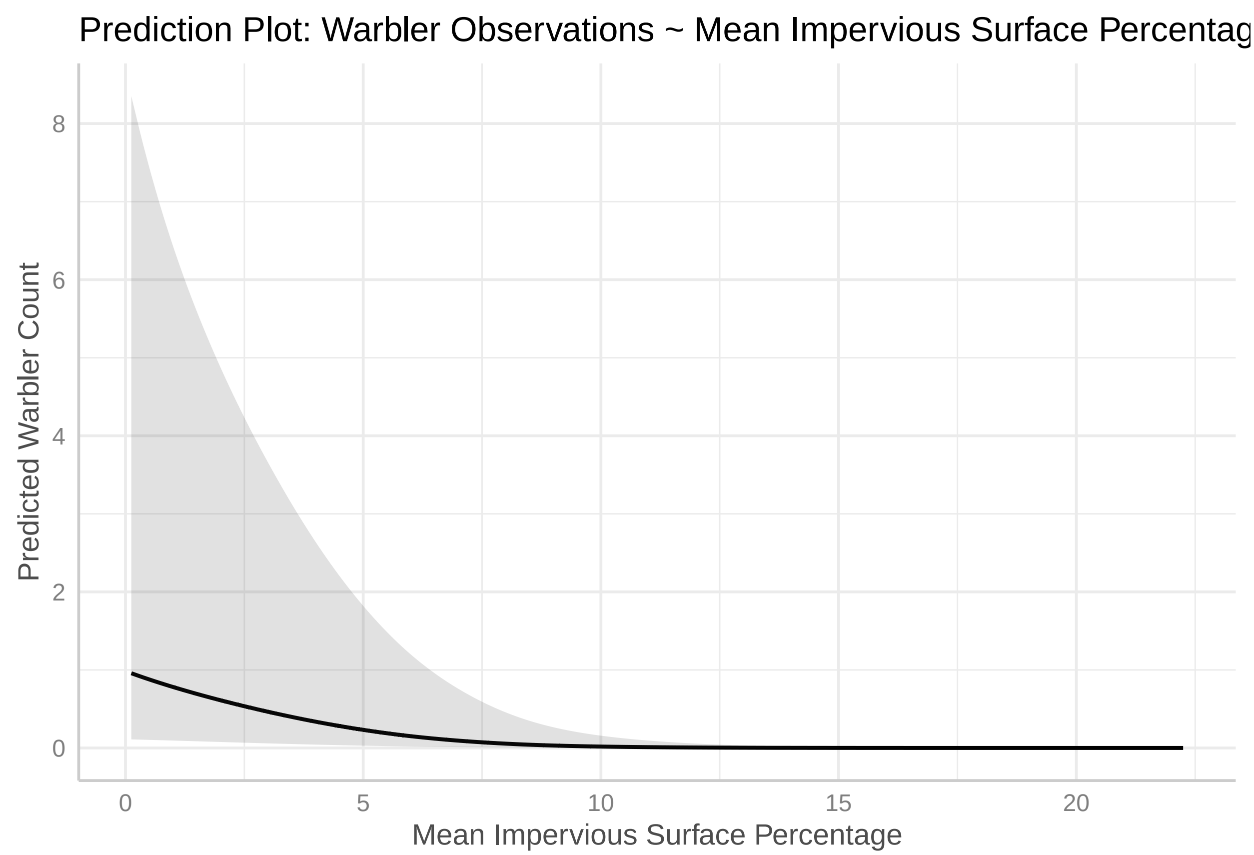
```
draw(bird_gam$gam) %>%
  gf_labs(x = "UTM E", y = "UTM N", title = "Contour of Warbler Observations")
```

```
## Warning: Removed 6464 rows containing non-finite values (stat_contour).
```

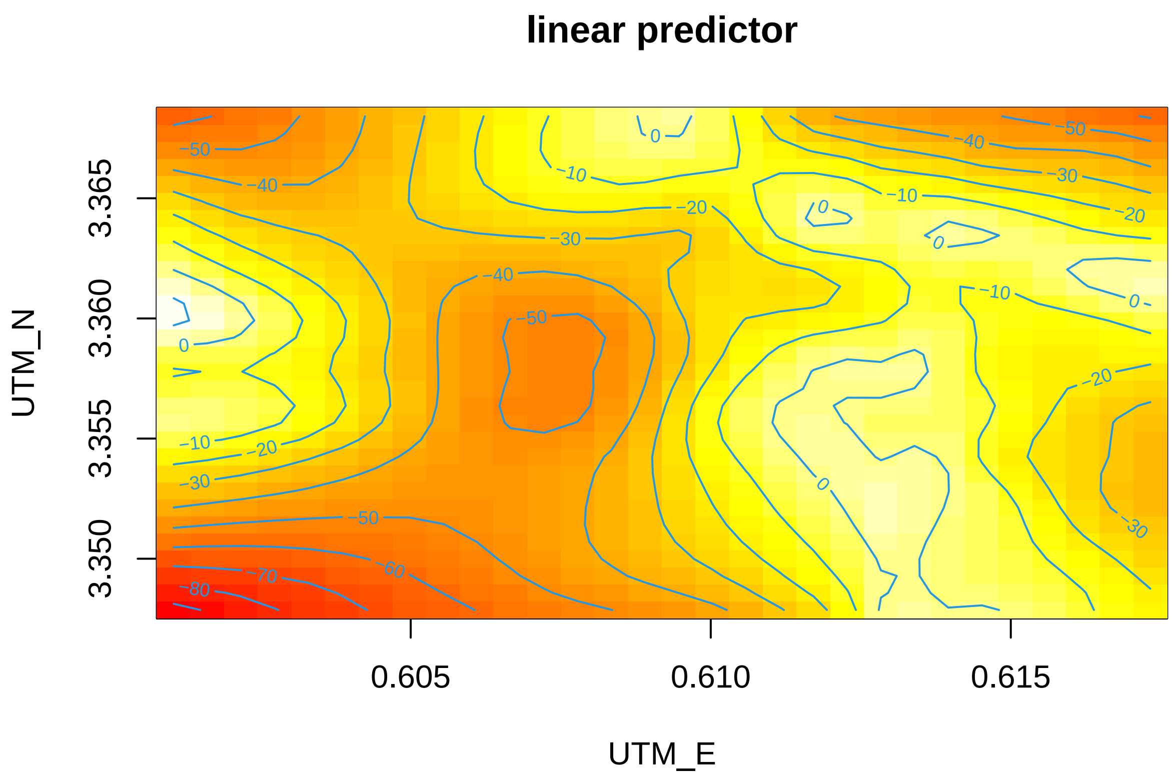


Prediction Plots

```
ggeffects::ggpredict(
  bird_gam,
  terms = c(
    "MeanImp"),
  type = "fixed"   ) %>%

  plot() %>%
  gf_labs(x = "Mean Impervious Surface Percentage", y = "Predicted Warbler Count", title = "Prediction
```




```r
vis.gam(bird_gam$gam, view = c("UTM_E", "UTM_N"), plot.type = "contour")
```



```
#birds %>%
  #filter(Year == 2011) %>%
  #ggplot(aes(x = UTM_E, y = UTM_N)) +
  #geom_tile(mapping = aes(fill = GCW))
```

Summary