HW2: Continuous, fixed spatial index

Fern Bromley

October 3, 2025

I. Mathematical [6 pts]

(1) [6 pts] From Section 6.6 in ZVH, do **either** exercise 5. or 9. For 9., three covariance functions from each table (6.1 and 6.2) is sufficient.

5. Consider an isotropic second-order stationary geostatistical process in $\mathbb{R}^d$ with covariance function

$$C(r; \sigma^2, \kappa) = \begin{cases} \sigma^2 & \text{if } r = 0, \\ \sigma^2 \kappa & \text{if } r \neq 0, \end{cases}$$

where $\sigma^2 > 0$. Show that κ cannot be negative. (Hint: Consider the positive-definiteness property (6.1), with $a_i = 1$ for every site i.)

- **Exercise 5:** A valid, parametric covariance function must be positive definite (i.e.,)

r = intersite distance; sigma2 = variance; variance times some constant if the site is further away. If $\sigma^2 > 0$, and the value of C must be 0, then κ must be > 0.

For the covariance function above to be valid,

II. Tucson Water [28 pts]

The Arizona Department Of Environmental Quality (ADEQ) monitors ground water for a large number of potentially hazardous chemicals at sites around the state. One such chemical is **1,4-Dioxane**, which has a number of industrial uses, but is also irritating to eyes and respiratory systems, and is a possible carcinogen. The data in `1_4_dioxane.csv` were gathered from **https://www.waterqualitydata.us/** and represent concentrations of the chemical 1,4-Dioxane in ground water near Tucson as measured in micrograms per liter (`ResultMeasureValue`). Each measurement is associated with a date (`AnalysisStartDate`) and the coordinates of the monitoring site (`Longitude`/`LatitudeMeasure`). In addition, a binary variable indicating whether or not the monitoring site is located within the boundary of Tucson International Airport (TIA) is also included (`airport`).

(2) [2 pts] Create a map like the one shown (see Gradescope PDF) to visualize the spatial arrangement of log-dioxane concentrations. Be sure to choose colors appropriate for the measured variable. **Provide executable code (PEC)**.

```
library(sf)
```

```
## Linking to GEOS 3.13.0, GDAL 3.8.5, PROJ 9.5.1; sf_use_s2() is TRUE
```

```
diox <- read.csv("1_4_dioxane.csv")
diox$log_diox <- log(diox$ResultMeasureValue)

library(ggmap)
```

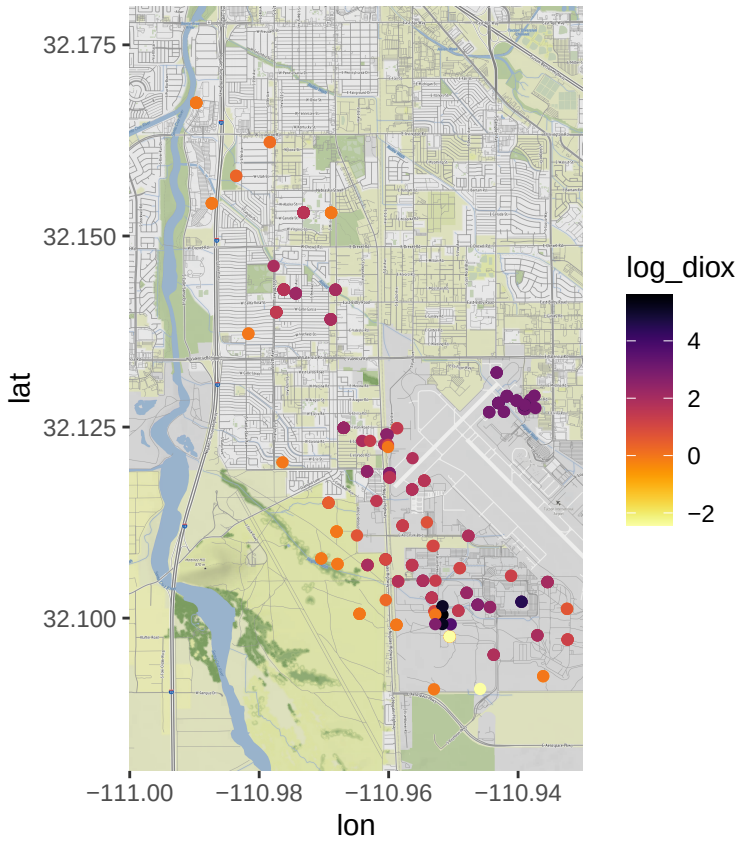
```
## Loading required package: ggplot2
```

```
## i Google's Terms of Service: <https://mapsplatform.google.com>
##   Stadia Maps' Terms of Service: <https://stadiamaps.com/terms-of-service>
##   OpenStreetMap's Tile Usage Policy: <https://operations.osmfoundation.org/policies/tiles>
## i Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```
bbox <- c(left = -111, bottom = 32.08, right = -110.93, top = 32.18)
diox_map <- ggmap::get_stadiamap(bbox = bbox, zoom = 15)
```

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
## i 84 tiles needed, this may take a while (try a smaller zoom?)
```

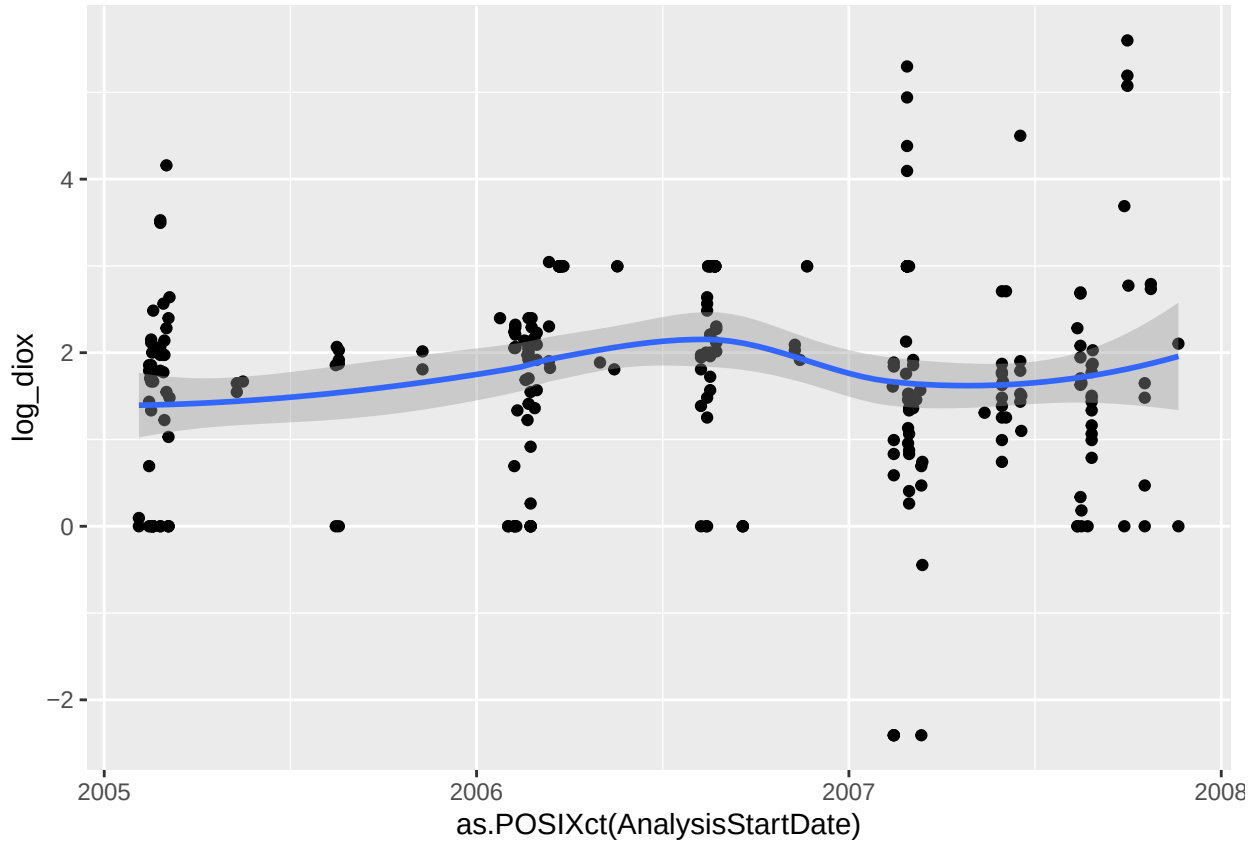
```
library(viridisLite)

ggmap(diox_map)+
  geom_point(data = diox,
             aes(x = LongitudeMeasure, y = LatitudeMeasure,
                 color = log_diox))+
  scale_color_gradientn(colors = viridis(256, option = "B", direction = -1))
```



(3) [3 pts] Make a scatterplot showing log-dioxane as a function of the date each measurement was taken. Does your plot suggest time is related to concentrations of dioxane? Make a figure with two boxplots of log-dioxane grouped by whether or not sites are located at the airport or not. Does your figure suggest that sites at the airport have meaningfully different concentrations than other sites?

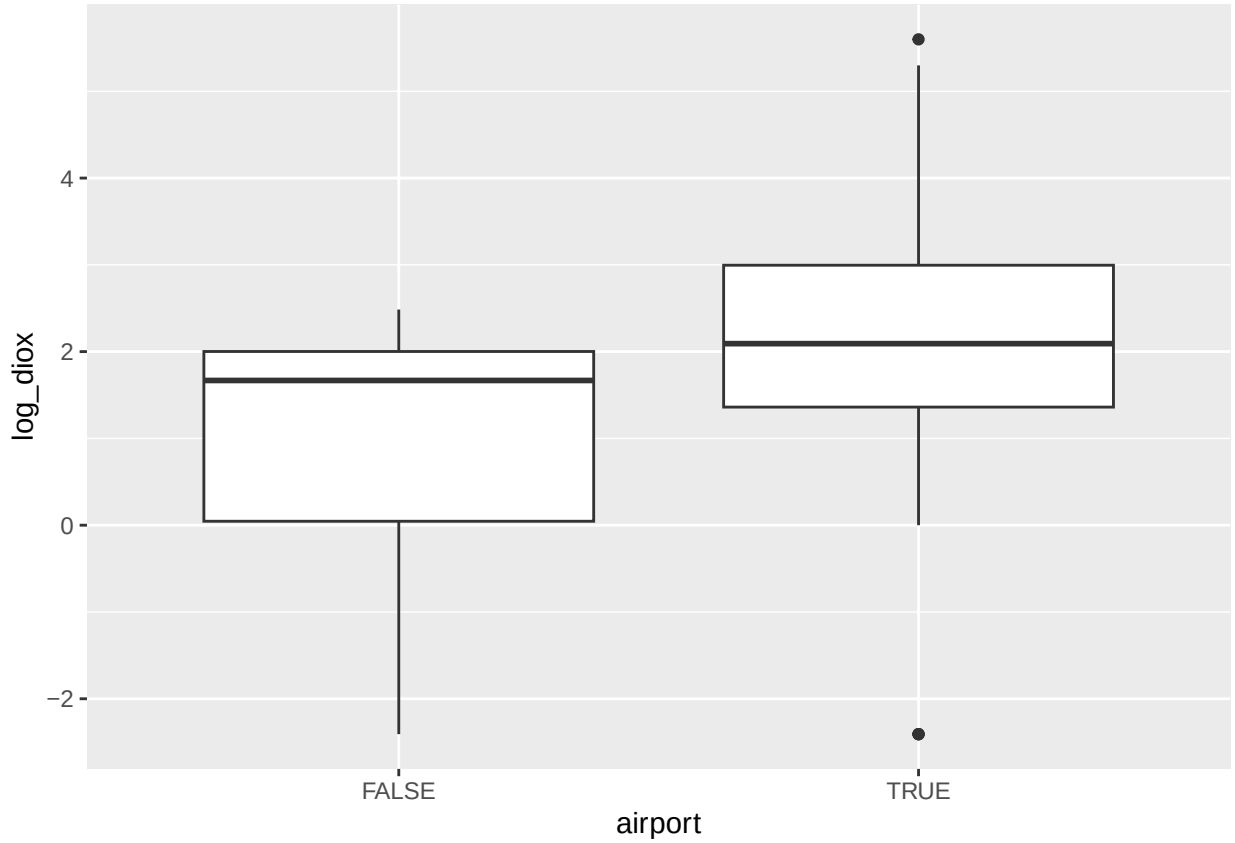
```
ggplot(diox, aes(x = as.POSIXct(AnalysisStartDate), y = log_diox))+
  geom_point()+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



It would not appear from the scatterplot that the log dioxane concentration is following a temporal pattern.

```
ggplot(diox)+
  geom_boxplot(aes(x = airport, y = log_diox))
```



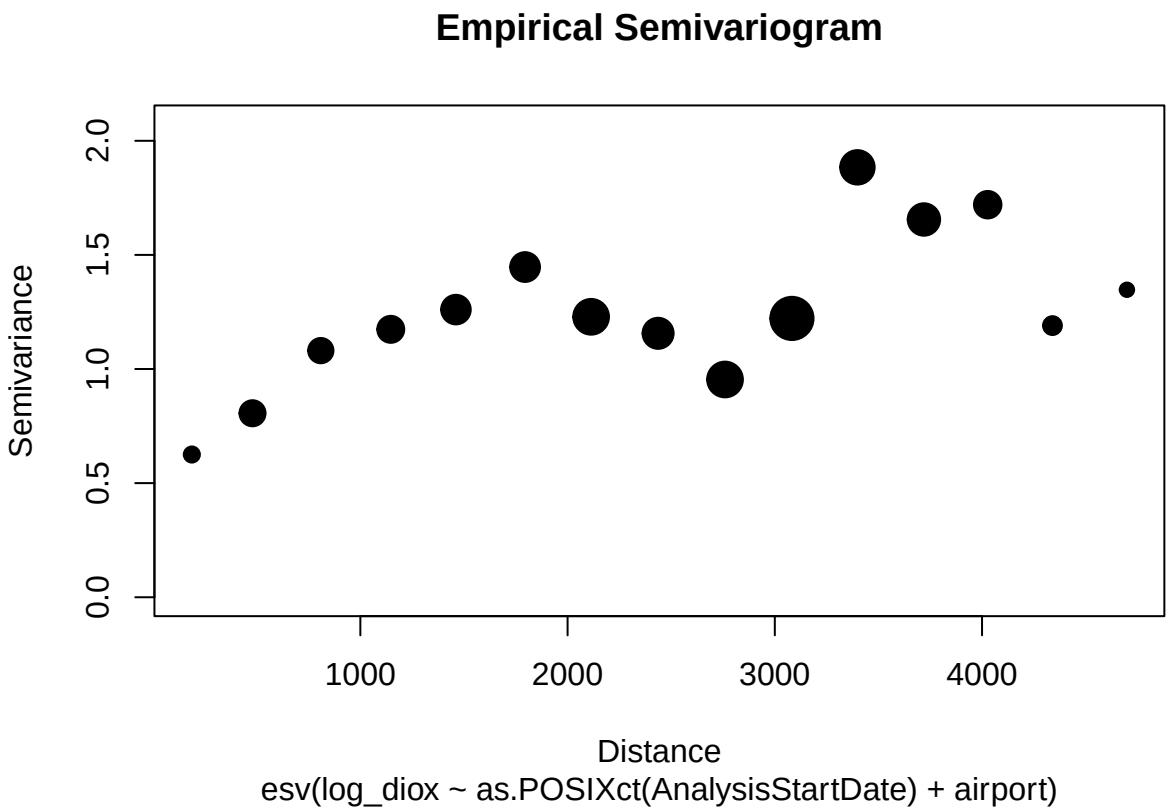
The box plot does suggest that the airport could have a higher mean log dioxane concentration.

(4) [3 pts] Transform the dioxane data to a new projection (coordinate reference system) corresponding to UTM zone 12N. **(PEC)**. Make an empirical semivariogram for the log-concentration of dioxane after accounting for possible linear effects of date and whether a site is at the airport or not. Give a rough estimate for the size of the nugget effect.

```
diox_sf <- st_as_sf(diox, coords = c("LongitudeMeasure", "LatitudeMeasure"),
                    crs = "epsg:4269")
diox_12n <- st_transform(diox_sf, crs = "epsg:32612")
```

```
# ok now we account for linear effect of date and the airport location:
library(spmodel)
plot(esv(log_diox ~ as.POSIXct(AnalysisStartDate) + airport, diox_12n))
```

```
## Warning: Zero distances observed between at least one pair. Ignoring pairs. If
## using splm(), consider a different estimation method.
```



The nugget effect estimated from the empirical semivariogram appears to be around 0.4 - 0.5.

(5) [3 pts] Fit a spatial linear regression model using restricted maximum likelihood (REML) to log-concentrations of dioxane as a linear function of measurement date and whether a site is located at TIA. Use the Matérn parametric family of covariance functions. Report the estimated nugget, partial sill, and range parameters. Given your semivariogram from the previous problem, do the parameter estimates make sense to you?

```
reml_mod <- splm(log_diox ~ as.POSIXct(AnalysisStartDate) + airport,
                 data = diox_12n, estmethod = "reml",
                 spcov_type = "matern")
reml_mod$coefficients
```

```
## $fixed
##                   (Intercept) as.POSIXct(AnalysisStartDate)
##                  5.491160e+00                 -3.944538e-09
##                   airportTRUE
##                  7.395120e-01
##
## $spcov
##          de          ie       range       extra      rotate       scale
##   1.6008280   0.1770568 101.5581745   3.1430398   0.0000000   1.0000000
## attr(,"class")
## [1] "matern"
##
## $randcov
## NULL
```

(6) [3 pts] Fit the same spatial linear regression model to the observations using the two-stage semivariogram + weighted least squares (SV-WLS) approach. Report the estimated covariance function parameters. Which estimation method, REML or SV-WLS, do you think yields the most reasonable covariance function parameters?

```
sv_wls_mod <- splm(log_diox ~ as.POSIXct(AnalysisStartDate) + airport,
                   data = diox_12n, estmethod = "sv-wls",
                   spcov_type = "matern")
```

```
## Warning: Zero distances observed between at least one pair. Ignoring pairs. If
## using splm(), consider a different estimation method.
```

```
sv_wls_mod$coefficients
```

```
## $fixed
##                    (Intercept) as.POSIXct(AnalysisStartDate)
##                   2.863819e+00                  -1.876329e-09
##                    airportTRUE
##                   1.744523e-01
##
## $spcov
##           de           ie        range        extra       rotate        scale
##    1.0203233    0.5425866 1380.3165398    0.4548983    0.0000000    1.0000000
## attr(,"class")
## [1] "matern"
##
## $randcov
## NULL
```

SV-WLS yielded more reasonable covariance function parameters.

(7) [3 pts] Use leave-one-out cross validation to compare the predictive performance of each fitted model (REML and SV-WLS). Which model is associated with the smallest mean squared prediction error?

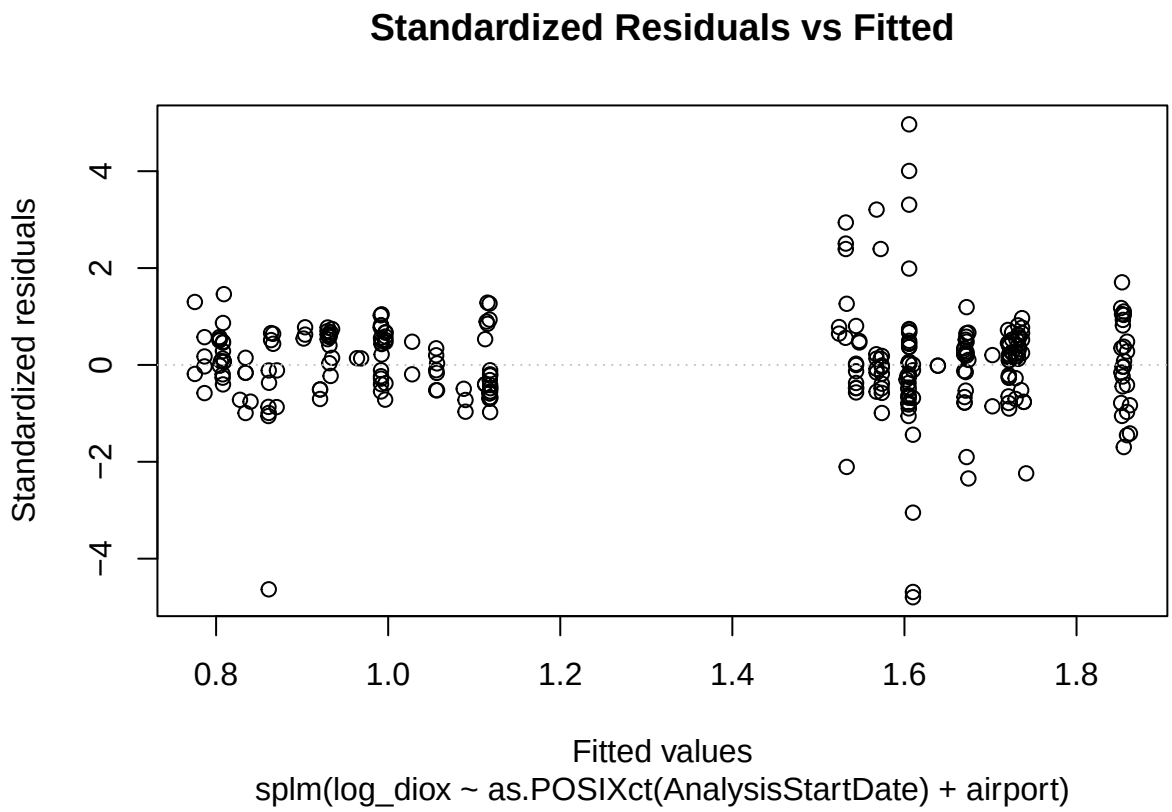
```
rbind(c("reml", spmodel::loocv(reml_mod)), c("sv-wls", spmodel::loocv(sv_wls_mod)))
```

```
##               bias         MSPE      RMSPE     cor2
## [1,] "reml"   0.001905365 0.3908703 0.6251962 0.735344
## [2,] "sv-wls" -0.01001001 0.4957254 0.7040777 0.6697993
```

The first model using REML estimation had a lower MSPE.

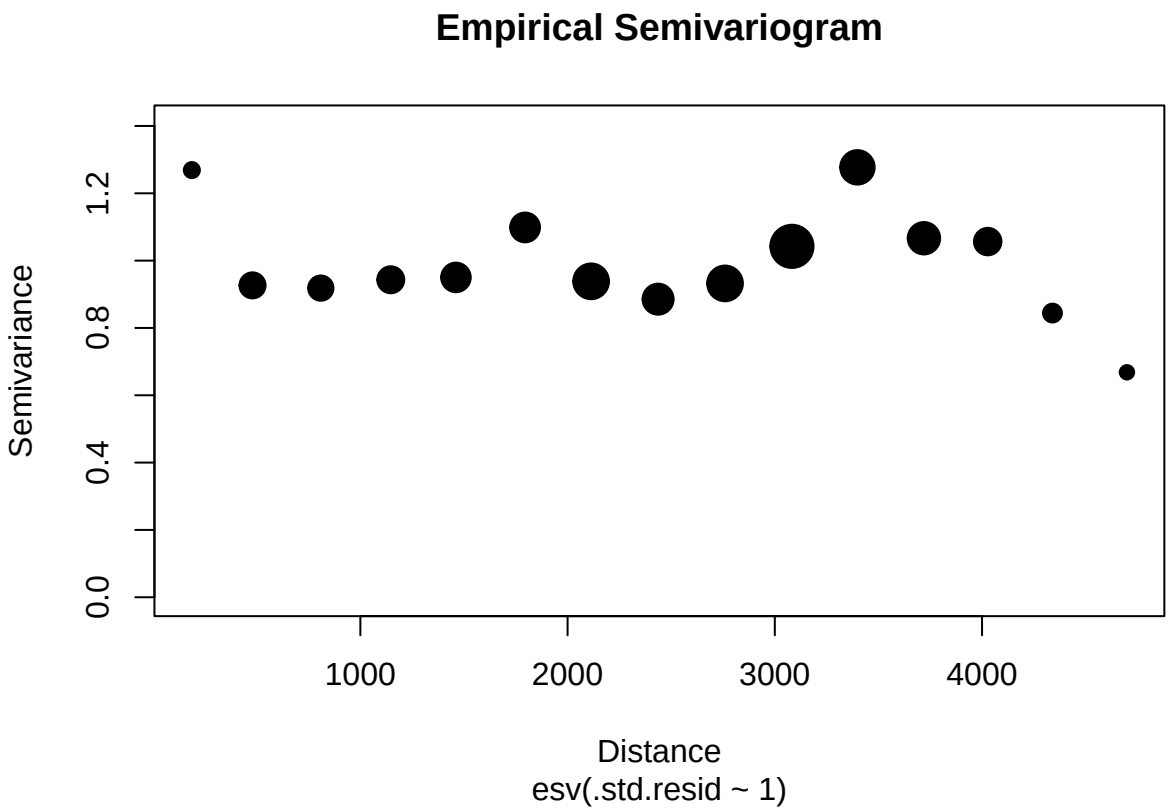
(8) [2 pts] Create diagnostic plots to visually assess how reasonable the assumption of marginal normality is for each fitted model. Interpret your plots.

```
plot(reml_mod, 1)
```

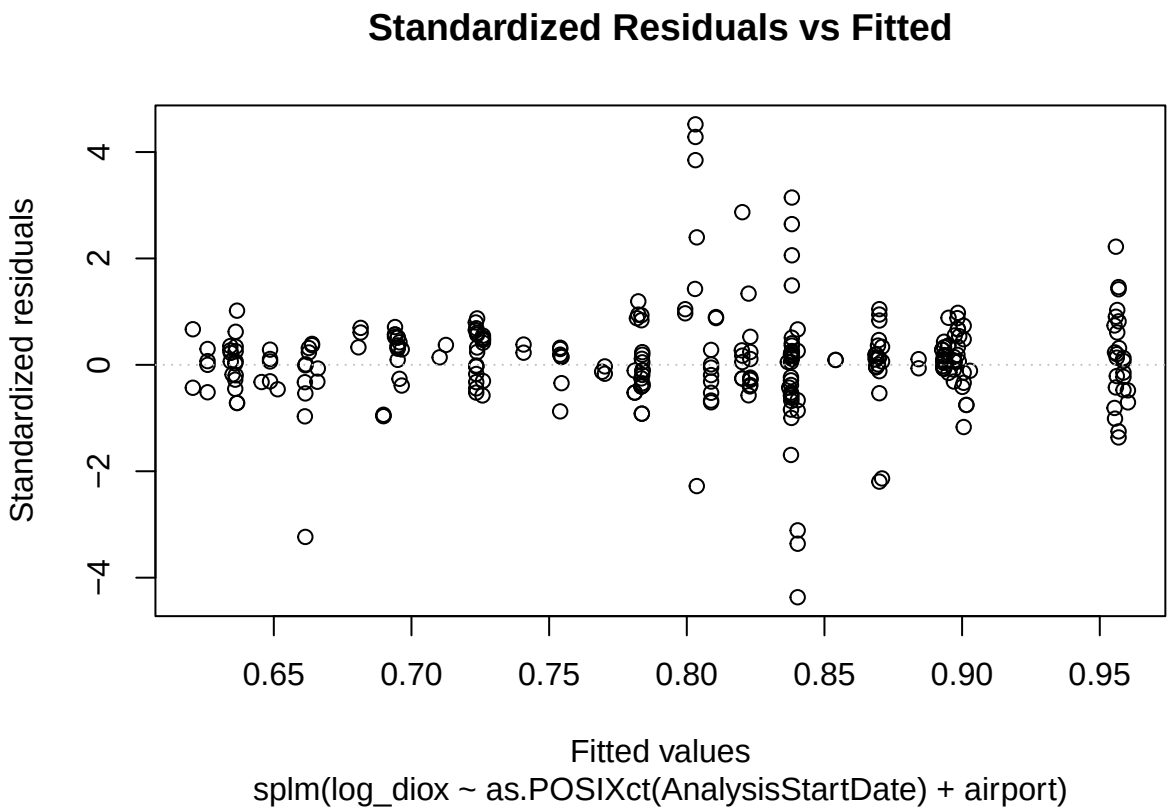



```
plot(esv(.std.resid ~ 1, augment(reml_mod)))
```

```
## Warning: Zero distances observed between at least one pair. Ignoring pairs. If
## using splm(), consider a different estimation method.
```

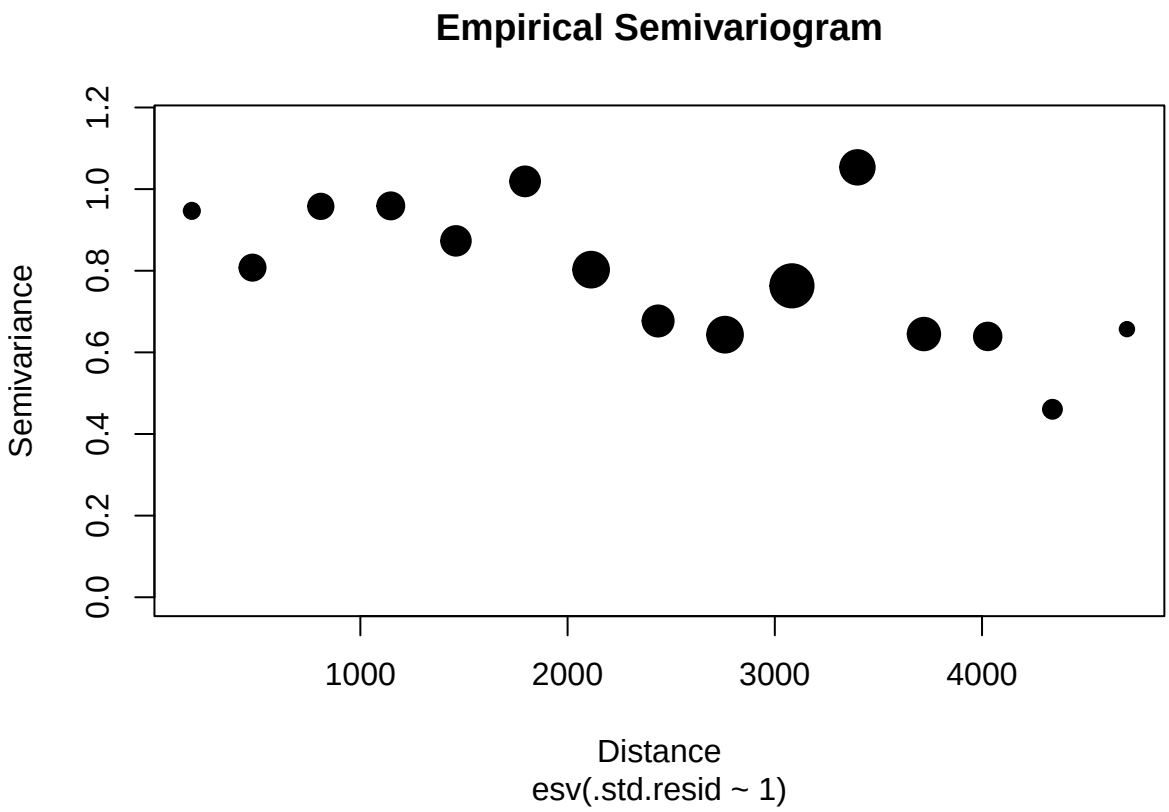



```
# hist(reml_mod$residuals$response, breaks = 100)

plot(sv_wls_mod, 1)
```




```
plot(esv(.std.resid ~ 1, augment(sv_wls_mod))) # adds model diagnostics as vars to the data (basically)
```

```
## Warning: Zero distances observed between at least one pair. Ignoring pairs. If
## using splm(), consider a different estimation method.
```




```
# hist(sv_wls_mod$residuals$response, breaks = 100)
```

For both models, the residuals appear normally distributed, and the standardized residuals do not show any spatial autocorrelation.

(9) [3 pts] Report and interpret the REML-estimated fixed effects of date and whether or not a site is at TIA. Do the signs match what you'd expect? Why/why not?

```
summary(reml_mod)
```

```
##
## Call:
## splm(formula = log_diox ~ as.POSIXct(AnalysisStartDate) + airport,
##     data = diox_12n, spcov_type = "matern", estmethod = "reml")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0178 -0.3806  0.5799  1.2714  4.0666
##
## Coefficients (fixed):
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    5.491e+00  1.271e+00   4.320 1.56e-05 ***
## as.POSIXct(AnalysisStartDate) -3.945e-09  1.080e-09  -3.652  0.00026 ***
## airportTRUE                    7.395e-01  3.233e-01   2.287  0.02218 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Pseudo R-squared: 0.06156
##
## Coefficients (matern spatial covariance):
##       de       ie    range    extra
##   1.6008   0.1771 101.5582   3.1430
```

The fixed effects estimates from the REML-fitted models suggest that there was a very small effect of time, such that as time passed, log dioxane concentrations decreased. Additionally, there was an effect of whether the site is at the airport or not, such that sites on the airport tended to have a lower log dioxane concentration. The negative coefficient for the effect of time makes sense if there are no longer inputs of dioxane into the environment and it is naturally being transported/decomposed, or if there are active dioxane cleanup efforts. I was surprised that sites at the airport were likely to have lower dioxane concentrations considering that it is an industrial chemical. However, perhaps the non-airport sites were also industrial as well.

(10) [2 pts] Use the REML-fitted model to create a 95% confidence interval for the expected log-concentration of dioxane in groundwater beneath the intersection of Drexel Rd. and 6th Ave. (32.1485, -110.9680) on May 28, 2007.

```
drexel <- data.frame(AnalysisStartDate = "2007-05-28", airport = F, LatitudeMeasure = 32.1485, Longitud
drexel_sf <- st_as_sf(drexel, crs = "epsg:32612", coords = c("LatitudeMeasure", "LongitudeMeasure"))

pred_diox <- predict(reml_mod, newdata = drexel_sf, se.fit = T) ; pred_diox
```

```
## $fit
##         1
## 0.8352816
##
## $se.fit
##        1
## 1.356951
```

```
pred_diox <- predict(reml_mod, newdata = drexel_sf, interval = "confidence"); pred_diox
```

```
##         fit       lwr      upr
## 1 0.8352816 0.3416563 1.328907
```

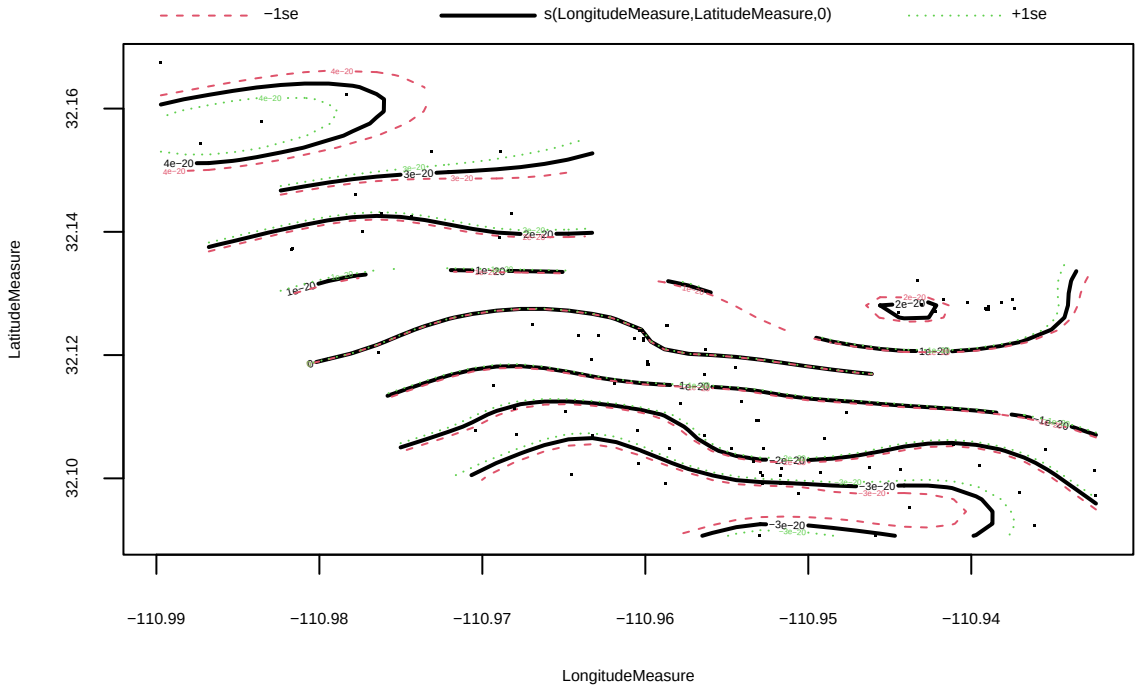
(11) [4 pts] Use a basis function approach to model the log-concentration of dioxane while accounting for the possible effects of date and whether or not a site is located at TIA. Use your fitted model to create another 95% confidence interval for the log-concentration of dioxane at Drexel Rd. and 6th Ave. on the same date. Which method produced the narrower confidence interval?

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
gam_mod <- gam(log_diox ~ as.POSIXct(AnalysisStartDate) + airport + s(LongitudeMeasure, LatitudeMeasure)
               data = diox)
plot(gam_mod)
```




```
pred_gam <- predict(gam_mod, newdata = drexel, se.fit = T)
pred_gam
```

```
## $fit
##        1
## 1.785145
##
## $se.fit
##          1
## 0.07366785
```

III. Canada Lynx [6 pts]

(12) [3 pts] Obtain the centered and scaled locations of two Canada lynx from the supplementary materials of **Buderman et al. (2016)**. Use the functionality of the `mgcv` package to fit independent GAM models to each coordinate of the bivariate location measurements of individual BC03F03. **Use cubic regression splines**, and experiment with the dimension of the basis (i.e., number of basis functions) to find a fit that looks good to you. **PEC**.

(13) [3 pts] Make two plots in the spirit of Figure 1(b) from Buderman et al. (2016) using your fitted models. Where in the two plots do you see the biggest discrepancies between your fit and the one from Buderman et al. (2016)?