HW2: Continuous, fixed spatial index

STAT 574E: Environmental Statistics

DUE: 10/4 11:59pm

I. Tucson Water

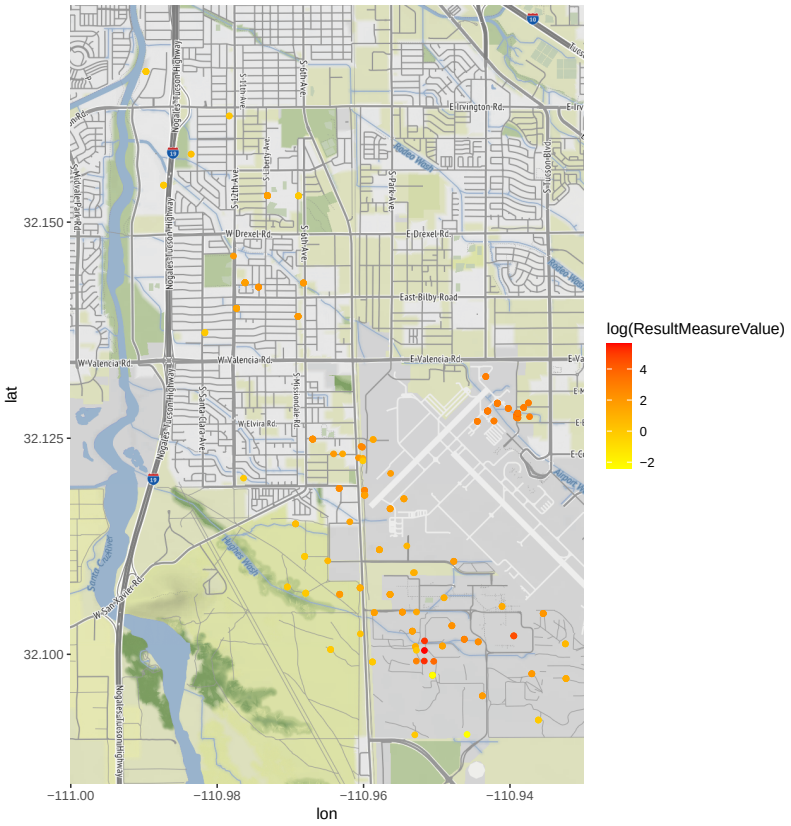
The Arizona Department Of Environmental Quality (ADEQ) monitors ground water for a large number of potentially hazardous chemicals at sites around the state. One such chemical is **1,4-Dioxane**, which has a number of industrial uses, but is also irritating to eyes and respiratory systems, and is a possible 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`).

(1) [3 pts] Create a map like the one shown to visualize the spatial arrangement of log-dioxane concentrations. Be sure to choose colors appropriate for the measured variable.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## 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.
## Linking to GEOS 3.11.2, GDAL 3.6.2, PROJ 9.2.0; sf_use_s2() is TRUE
##
## Rows: 285 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): MonitoringLocationIdentifier, CharacteristicName, ResultMeasure.Me...
## dbl  (3): ResultMeasureValue, LongitudeMeasure, LatitudeMeasure
## lgl  (1): airport
## date (1): AnalysisStartDate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## [1] "Date"
```

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```



(2) [4 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?

(3) [3 pts] Transform the dioxane data so that their new projection (coordinate reference system) corresponds 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.

(4) [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?

(5) [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?

(6) [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?

(7) [3 pts] Create diagnostic plots to visually assess how reasonable the assumption of marginal normality is for each fitted model. Interpret your plots.

(8) [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?

(9) [3 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.

(10) [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?

II. Canada Lynx [8 pts]

(11) [4 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**.

(12) [4 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)?