Homework 3, Generalized Linear Mixed Models

Jessica Chau

2019-04-09

Non-parametrics

Introduction

Carbon dioxide concentrates have risen compared to the past decades. This event is mainly caused by increased carbon dioxide emissions from the utilization of fossil fuels, including the user of vehicles. Carbon dioxide is a greenhouse gas, which absorbs sunlight and releases heat gradually to warm the Earth. This is the primary reason that the Earth's average temperate is 60 degrees instead of below freezing. However, higher levels of carbon dioxide signifies that the Earth's average temperate will continue to rise, causing global warming. It's important to be aware of carbon dioxide levels because of their ties to global warming and it's impact on extreme weather events, rising seas and the shifting of the population and habitats.

In this study, we will dive into historical Hawaiian carbon dioxide levels and answer the following hypotheses:

- Although carbon in the atmosphere is still increasing, there are indications that the increase has slowed somewhat recently.
- The data are consistent with carbon slowing during the global economic recessions around 1980-1982 and 2008, and the collapse of the Soviet Union after 1989
- Carbon tends to be higher in October than March
- There is a reasonable chance that carbon will exceed 430 parts per gallon by 2025

Figure 1 displays the carbon levels in Hawaii from 1960 March to 2018 October.

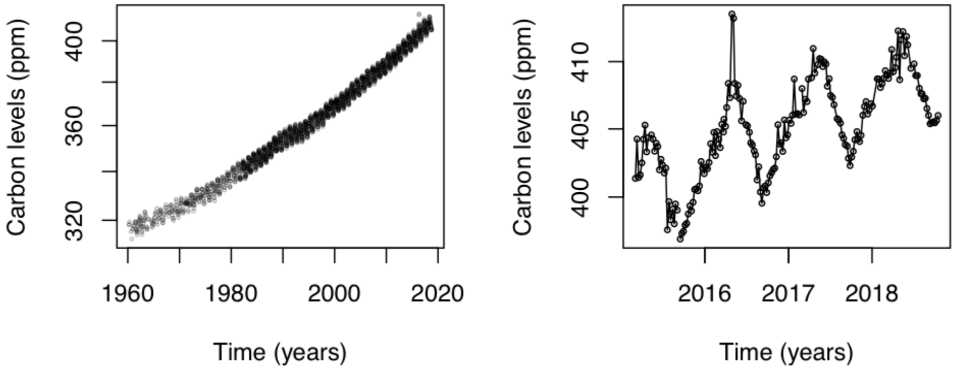

Figure 1: Carbon Levels at Mauna Los Observatory, Hawaii

Methods

The data in this study contains daily carbon dioxide levels. To analyze the data, a Generalized Additive Model (GAM) is used to model the data. A generalized additive model is used because it takes into consideration the nonparametric parameter f(s) which is a smoothing-varying function of days (as shown in equation 1).

$$Y_{ijk} \sim N(\lambda_i, \tau^2)$$

$$\begin{aligned}\lambda_i = &\beta_1 \cos(2\pi days/365.25) + \beta_2 \sin(2\pi days/365.25) \\ &+ \beta_3 \cos(4\pi days/365.25) + \beta_4 \sin(4\pi days/365.25) \\ &+ f(days; v)\end{aligned} \tag{1}$$

A check of the residuals in Figure 2 shows that the model fits the data well because the residuals are normally distributed.

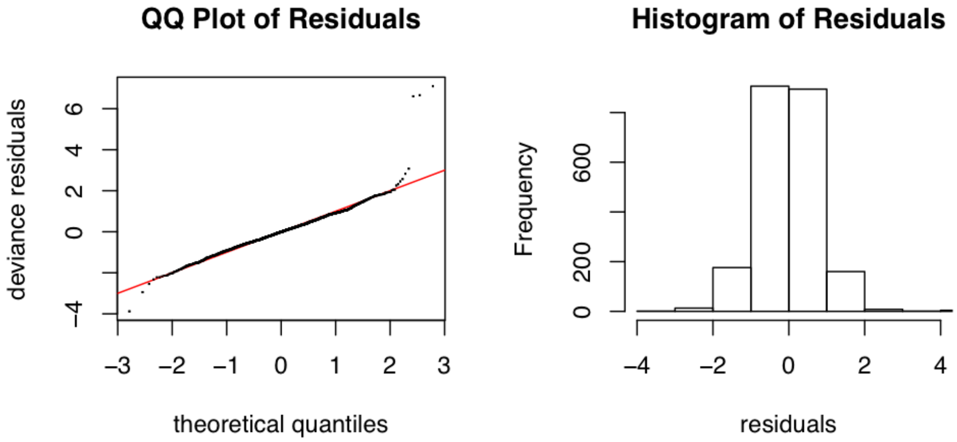

Figure 2: Generalized Additive Model Residuals Assessment

Table 1 shows the results from the GAM model, the covariates explain the seasonality trends.

To prove whether cardon dioxide levels have slowed recently (2018), the first derivative is calculated and plotted in Figure 3. A flat line indicates that there is no change in the acceleration of carbon dioxide, a positive slope indicates increasing acceleration levels of carbon dioxide and a negative slope indicates slower acceleration.

To confirm if carbon levels slowed during recessions, the same process is applied to determine if carbon levels slowed recently - by taking the first derivatives. Figure 3 shows the slope of carbon dioxide levels.

Table 1: Summary of GAM Results

	Estimate	Std. Error
(Intercept)	364.8868	0.01712
sin12	2.8652	0.02401
cos12	-0.9002	0.02442
sin6	-0.6078	0.02414
cos6	0.6449	0.02429

To determine if carbon levels are higher in October than in March, we need to take into consideration the number of days. Since October and March both have 31 days, we do not need to generate an offset term. The average carbon levels in March and October since 1980 are plotted as shown in Figure 4.

The last hypothesis requires predicting whether carbon will exceed 430 gallons. The GAM model is used to extrapolate the data up until 2030 and confidence intervals are plotted.

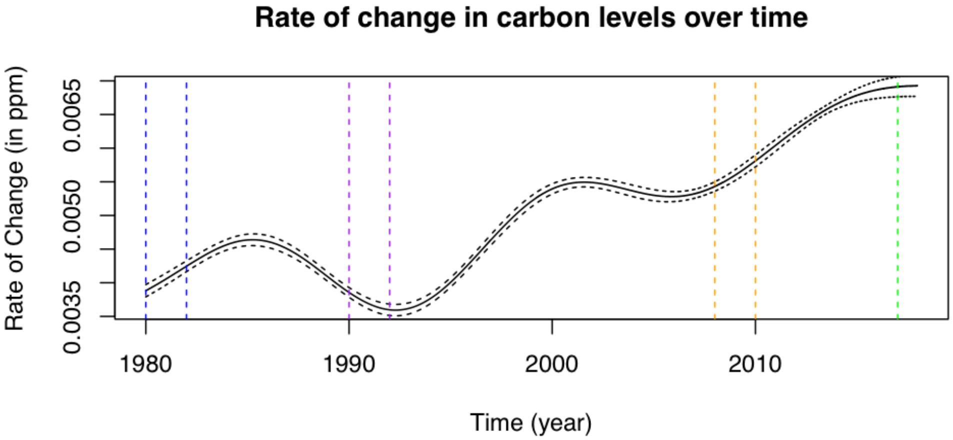

Figure 3: Assessment of carbon levels speed of change

Results

Hypothesis 1:

As shown on in Figure 3, there is a positive slope at the green vertical line that represents recent years (2017). This indicates that in recent months, carbon dioxide levels has not slowed down.

Hypothesis 2:

From Figure 3, the blue lines indicate the interval of the global economic recession (1980-1982). The upward slope indicates that the rate of change in carbon dioxide levels did not slow down during the recession. The purple lines indicate the collapse of the Soviet Union (1990-1992). The graph shows that carbon levels slowed down during this period. The orange lines indicate the recession in 2008 and carbon levels increased. The data shows that global economic recessions do not decrease carbon levels. The collapse of the Soviet Union did show slowing of carbon levels.

Hypothesis 3:

Figure 4 shows carbon levels per month in 2016 (the purple lines indicate March 1-31 and the green lines indicate October 1-31). Since seasonality of months is cyclical, the year 2016 is selected randomly and plotted in Figure 4. This shows that carbon levels in October do not exceed March.

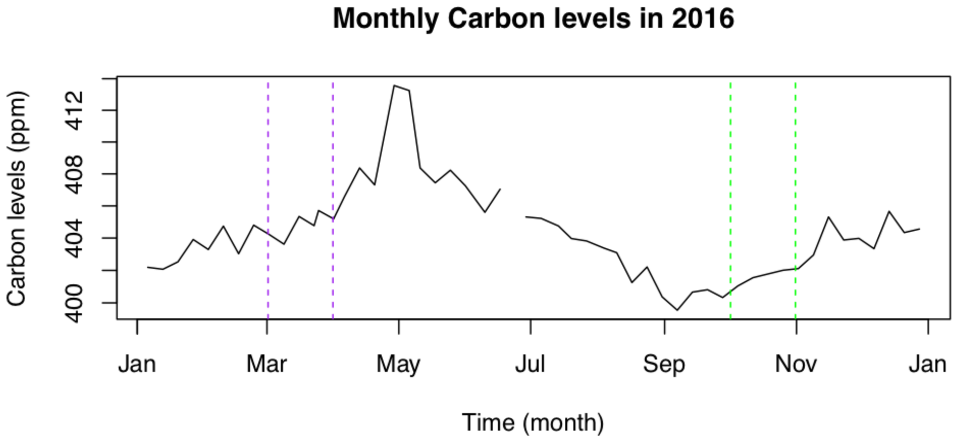

Figure 4: Comparing March to October Carbon Levels

Hypothesis 4:

From Figure 5, it shows the prediction of carbon levels until 2030. The purple vertical lines indicates the beginning and end of 2025 and the green horizontal line indicates carbon levels at 430ppm. The estimate carbon levels does not appear to exceed 430ppm by 2025, however the 95% upper confidence interval does appear to reach 430ppm.

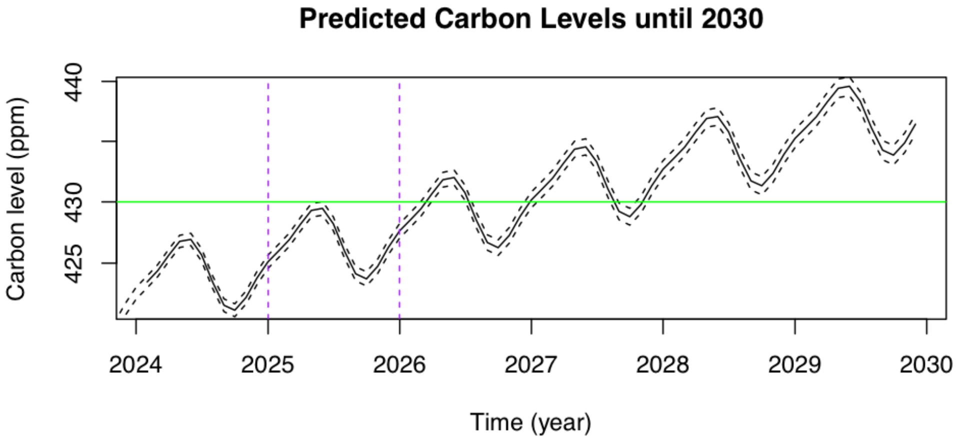

Figure 5: Predicted carbon levels until 2030

Discussion

Figure 3 shows that recessions (1980 & 2008) did not have an impact on carbon levels but the collapse of the Soviet Union did impact carbon levels. Figure 3 also shows that in recent periods (2017), carbon levels have not slowed down. This indicates that more action is needed to keep carbon levels from rising uncontrollably.

Figure 4 shows that carbon levels in October do not exceed March, as a result, hypothesis 3 is rejected.

As shown in Figure 5, carbon levels do not exceed 430ppm by January 1, 2025, however there is a chance that within 2025 carbon levels will be at 430ppm (by the 95% confidence interval).

With the conclusions on the hypothesis, it is important for the government of Hawaii to be viligant of the rising carbon levels and to keep it within the appropriate levels.

Malaria

Introduction

Malaria is a serious disease can will lead to coma or death if left untreated. Malaria cannot be transmitted from physical contact but is transmitted through mosquitoes. Due to this transmission method, it is common for children within close proximity to be infected because they live in an area that is active with Malaria. In order to efficiently provide care to children, it is important to understand which regions are highly likely for Malaria transmission to occur. This study evaluates whether it is feasible to predict the spatial distribution of malaria in Gambia.

Figure 6 shows the map of Gambia and the locations where samples were taken from are marked by a black '+'. This shows that samples were taken in 4 major areas of Gambia.

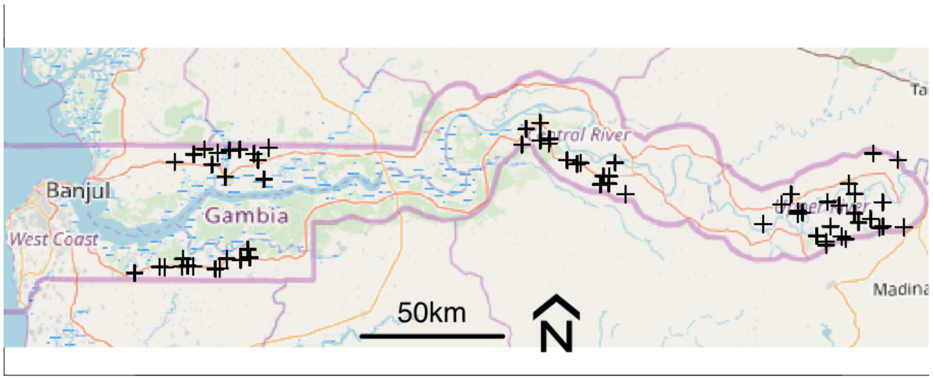

Figure 6: Samples taken of Children in Gambia

Methods

A geospatial model is built to evaluate the spatial distribution of Malaria. A generalized linear geostatistical model (GLGM) is selected because the outcome of the data is binary (0 or 1) and suitable for logistic regression. The $U(s)$ is the residual spatial variation, and it is the difference between actual malaria concentration and what the covariates predict and it depends on σ (variability in residual), ϕ (range parameter) and v (shape parameter).

$$Y_i \sim Binomial(\lambda(s_i))$$
$$logit(\lambda(s_i)) = \mu + \beta_1(evi) + \beta_2(phc) + U(s)$$
$$cov[U(s+h), U(s)] = \sigma^2 \rho(\frac{h}{\phi}; v)$$

The model has two covariates: evi (vegetation) and pch (whether the village has a public health centre). The model requires specification of the priors for σ and ϕ. The selected range is 65km because according to the American Mosquito Control Association, the average mosquito travels 35-65km without a host. As mentioned in the introduction, it is difficult for malaria to be transmitted through contact between humans. Transmission is mainly through areas with a high density mosquitos. The prior for σ controls for the variabilty in residual variation, which should be small; there should be small residual spatial variation.

Figure 7 displays the prior and posterior distributions of σ and ϕ. The posterior of sd and range is highest at 1 and 30 respectively. The posterior curves are not flat; the model appears to be well defined and analyses can be built from the model.

Table 2: Summary of GLGM Results

	mean	0.025quant	0.975quant	meanExp
Intercept	-0.83760	-2.21866	0.63025	0.5484
Public Health Centre	-0.34026	-0.69429	0.01836	0.7187
Vegetation Index	0.00037	-0.00029	0.00102	1.0040
range/1000	30.05075	15.54647	57.22987	NA
sd	1.06065	0.72600	1.62748	NA

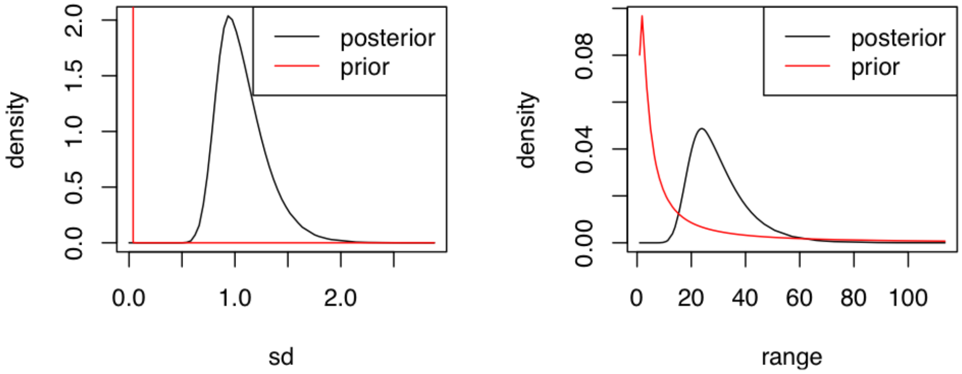

Figure 7: Posteriors and prior comparisons of sd and range parameters

Results

Table 2, shows the log odds of the results from the model. The odds of having a public health center in the village is 0.7116 (calculated by exp(-0.34026)), indicating that having a public health centre reduces the odds of having Malaria. The odds of having Malaria is slightly higher in areas with a higher vegatation index.

The residuals and correlations are plotted with respect to the samples as shown in Figure 8 and 9 respectively. The correlation on Gambia's east region is more positively correlation (the red) compared to the other regions in Gambia as shown in Figure 8 The residual spatial variation shown in Figure 9 indicates that the likelihood of Malaria is higher for the eastern region of Gambia.

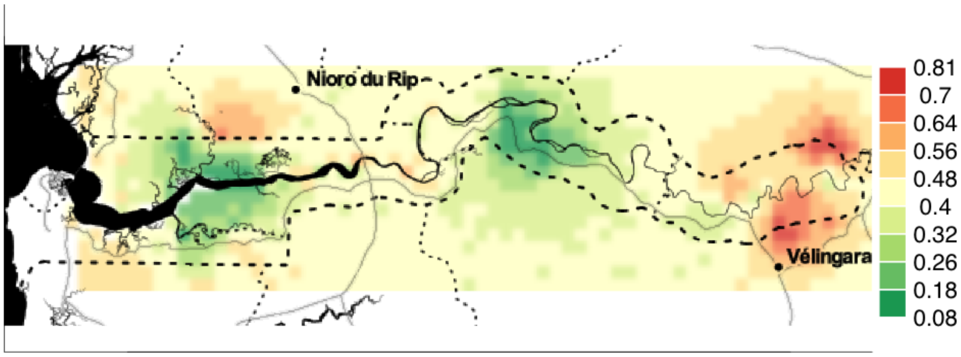

Figure 8: Posterior correlation for spatial surfaces calculated from glgm model

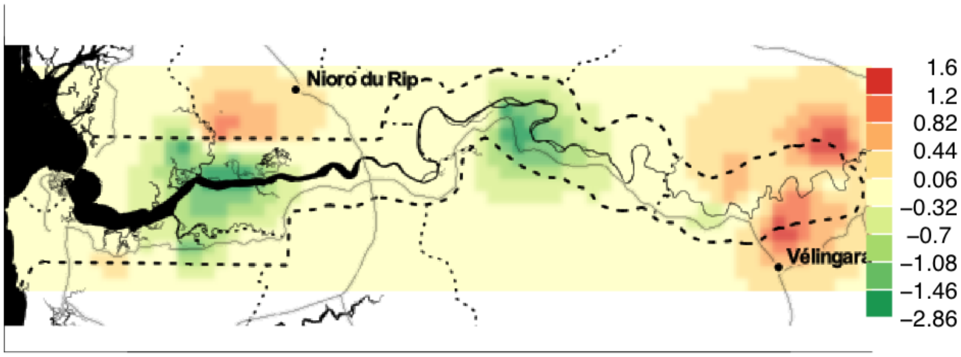

Figure 9: Posterior means for spatial surfaces calculated from glgm model

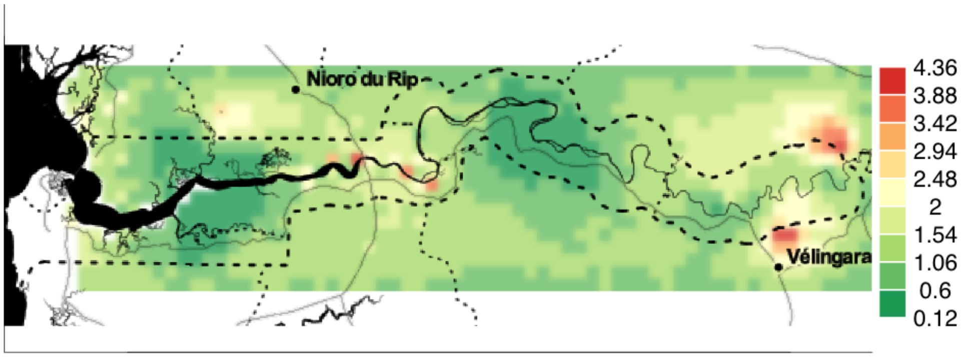

Figure 10: Predicted Malaria in Children from glgm

Discussion

From the generalized linear geostatistical model on the Malaria dataset, it shows that Malaria is prevalent in the eastern regions and Figure 10 further supports this. The GLGM was an adequate model on predictions of Malaria transmissions because it considers the covariance in surrounding regions which other statistical models cannot perform.

References

Centers for Disease Control and Prevention. "About Malaria: Frequently Asked Questions." CGC.gov. https://www.cdc.gov/malaria/about/faqs.html (accessed April 7, 2019)

American Mosquito Control Association. "Frequently Asked Questions." Mosquito.org. https://www.mosquito.org/page/FAQ (accessed April 7, 2019)

R Code Reference

```r
cUrl = paste0("http://scrippsco2.ucsd.edu/assets/data/atmospheric/",
"stations/flask_co2/daily/daily_flask_co2_mlo.csv")
cFile = basename(cUrl)
if (!file.exists(cFile)) download.file(cUrl, cFile)
co2s = read.table(cFile, header = FALSE, sep = ",", skip = 69,
stringsAsFactors = FALSE, col.names = c("day", "time",
"junk1", "junk2", "Nflasks", "quality", "co2"))
co2s$date = strptime(paste(co2s$day, co2s$time), format = "%Y-%m-%d %H:%M", tz = "UTC")
# remove low-quality measurements
co2s[co2s$quality > 2, "co2"] = NA

# par(mfrow=c(1,2))
# plot(co2s$date, co2s$co2, log = "y", cex = 0.3, col = "#00000040",
# xlab = "Time (years)", ylab = "Carbon levels (ppm)")
# plot(co2s[co2s$date > ISOdate(2015, 3, 1, tz = "UTC"), c("date",
# "co2")], log = "y", type = "o", xlab = "Time (years)", ylab = "Carbon levels (ppm)",
# cex = 0.5)


timeBreaks = seq(min(co2s$date), ISOdate(2025, 1, 1, tz = "UTC"), by = "14 days")
timePoints = timeBreaks[-1]

timeOrigin = ISOdate(1980, 1, 1, 0, 0, 0, tz = "UTC")
co2s$days = as.numeric(difftime(co2s$date, timeOrigin, units = "days"))
co2s$cos12 = cos(2 * pi * co2s$days/365.25)
co2s$sin12 = sin(2 * pi * co2s$days/365.25)
co2s$cos6 = cos(2 * 2 * pi * co2s$days/365.25)
co2s$sin6 = sin(2 * 2 * pi * co2s$days/365.25)
# cLm = lm(co2 ~ days + cos12 + sin12 + cos6 + sin6, data = co2s)
# tbl_co <- summary(cLm)$coef[, 1:2]

### MY CODE ###
# https://stats.stackexchange.com/questions/380426/when-to-use-a-gam-vs-glm
# https://kevintshoemaker.github.io/NRES-746/GAMs.html#additive_models
co2s$month = months(as.Date(co2s$day))
gam_mdl = gam(co2 ~ sin12 + cos12 + sin6 + cos6 + s(days), data=co2s)
# gam_mdl$sp

# gam.check(gam_mdl) #Has a good fit

# plot(gam_mdl,pages=1)
# gam.check(gam_mdl)
#
# qq.gam(gam_mdl)
# qq.gam(gam_mdl)

# par(mfrow=c(1,2))
# type <- "deviance"
# resid <- residuals(gam_mdl, type = type)
# qq.gam(gam_mdl, rep = 0, level = 0.9, type = type, rl.col = 2,
#        rep.col = "gray80", main="QQ Plot of Residuals")
# hist(resid, xlab = "residuals", main = "Histogram of Residuals", xlim=c(-4,4))


tbl_co <- summary(gam_mdl)
timeBreaks = seq(min(co2s$date), ISOdate(2025, 1, 1, tz = "UTC"), by = "14 days")
timePoints = timeBreaks[-1]

timeOrigin = ISOdate(1980, 1, 1, 0, 0, 0, tz = "UTC")
co2s$days = as.numeric(difftime(co2s$date, timeOrigin, units = "days"))
co2s$cos12 = cos(2 * pi * co2s$days/365.25)
co2s$sin12 = sin(2 * pi * co2s$days/365.25)
co2s$cos6 = cos(2 * 2 * pi * co2s$days/365.25)
co2s$sin6 = sin(2 * 2 * pi * co2s$days/365.25)
# cLm = lm(co2 ~ days + cos12 + sin12 + cos6 + sin6, data = co2s)
# tbl_co <- summary(cLm)$coef[, 1:2]



### MY CODE ###
# https://stats.stackexchange.com/questions/380426/when-to-use-a-gam-vs-glm
# https://kevintshoemaker.github.io/NRES-746/GAMs.html#additive_models
co2s$month = months(as.Date(co2s$day))
gam_mdl = gam(co2 ~ sin12 + cos12 + sin6 + cos6 + s(days), data=co2s)
# gam_mdl$sp
# knitr::kable(summary(gam_mdl)$p.table[,1:2],digits=5, escape = FALSE, format = "latex",
#               caption = 'Summary of GAM Results')


#hypothesis 1 & 2
newX = data.frame(date = seq(ISOdate(1980, 1, 1, 0, 0, 0,
tz = "UTC"), by = "1 days", length.out = 365 * 38))
base_x = as.numeric(difftime(newX$date, timeOrigin, units = "days"))
len_v = length(base_x)
newdf = data.frame(days=base_x, sin12=integer(len_v), cos12=integer(len_v), sin6=integer(len_v
delta = 1e-7 # finite difference interval

base_x_delta = base_x + delta ## shift the evaluation
newdf_delta = data.frame(days=base_x_delta, sin12=integer(len_v), cos12=integer(len_v), sin6=i

x_predict = predict(gam_mdl, newdf, type="lpmatrix")
x_predict_delta = predict(gam_mdl, newdf_delta, type="lpmatrix")
first_derivative = (x_predict_delta-x_predict)/delta
Xp = (x_predict_delta-x_predict)/delta
# colnames(Xp)
# head(first_derivative)

# dim(first_derivative)
# length(newX$date)
# plot(first_derivative) this works

# i=1
# Xi = Xp*0
# Xi[,6:14] <- Xp[,6:14]
# df = Xi%*%coef(gam_mdl)
# df.sd = rowSums(Xi%*%gam_mdl$Vp*Xi)^.5
# plot(newX$date, df, type = "l", xlab = 'Time (year)',
#      ylab='Rate of Change (in ppm)', main='Rate of change in carbon levels over time')
# lines(newX$date, df+2*df.sd,lty=2)
# lines(newX$date, df-2*df.sd,lty=2)
# abline(v = ISOdate(2017, 1, 1, tz = "UTC"), col="green", lty=2)
# abline(v = ISOdate(1980, 1, 1, tz = "UTC"), col="blue", lty=2)
# abline(v = ISOdate(1982, 1, 1, tz = "UTC"), col="blue", lty=2)
# abline(v = ISOdate(1990, 1, 1, tz = "UTC"), col="purple", lty=2)
# abline(v = ISOdate(1992, 1, 1, tz = "UTC"), col="purple", lty=2)
# abline(v = ISOdate(2008, 1, 1, tz = "UTC"), col="orange", lty=2)
# abline(v = ISOdate(2010, 1, 1, tz = "UTC"), col="orange", lty=2)


# Testing out different models
# gam_mdl2 = gam(co2 ~ sin12 + cos12 + sin6 + cos6 + month + s(days), data=co2s)
gam_mdl3 = gam(co2 ~ month + s(days), data=co2s)
```

```
# summary(gam_mdl3)
# gam.check(gam_mdl3)

# plot(co2s[(co2s$date >= ISOdate(2016, 1, 1, tz = "UTC") & co2s$date <= ISOdate(2016, 12, 31,
# "co2")], log = "y", type = "l", xlab = "Time (month)", ylab = "Carbon levels (ppm)",
# cex = 0.5, main="Monthly Carbon levels in 2016")
# abline(v = ISOdate(2016, 3, 1, tz = "UTC"), col="purple", lty=2)
# abline(v = ISOdate(2016, 3, 31, tz = "UTC"), col="purple", lty=2)
# abline(v = ISOdate(2016, 10, 1, tz = "UTC"), col="green", lty=2)
# abline(v = ISOdate(2016, 10, 31, tz = "UTC"), col="green", lty=2)

# Hypothesis 4: There is a reasonable chance that carbon will exceed 430 parts per gallon by 2
newX = data.frame(date=seq(from=timeOrigin, by="months", length.out=12*50))
# newX = data.frame(date = seq(ISOdate(1980, 1, 1, 0, 0, 0, tz = "UTC"), by = "1 days", length
newX$days = as.numeric(difftime(newX$date, timeOrigin, units = "days"))
newX$cos12 = cos(2 * pi * newX$days/365.25)
newX$sin12 = sin(2 * pi * newX$days/365.25)
newX$cos6 = cos(2 * 2 * pi * newX$days/365.25)
newX$sin6 = sin(2 * 2 * pi * newX$days/365.25)
newX$year = year(newX$date)

co2Pred = predict(gam_mdl, newX, se.fit = TRUE)
co2Pred = cbind(newX, co2Pred)
#head(co2Pred)
#
# plot(co2Pred[co2Pred$date >= ISOdate(2024, 1, 1, tz = "UTC"), c("date",
# "fit")], log = "y", type = "l", xlab = "Time (year)", ylab = "Carbon level (ppm)",
# cex = 0.5, main="Predicted Carbon Levels until 2030")
# lines(co2Pred$date, co2Pred$fit+2*co2Pred$se.fit,lty=2)
# lines(co2Pred$date, co2Pred$fit-2*co2Pred$se.fit,lty=2)
# abline(v = ISOdate(2025, 1, 1, tz = "UTC"), col="purple", lty=2)
# abline(v = ISOdate(2025, 12, 31, tz = "UTC"), col="purple", lty=2)
# abline(h = 430, col="green")



# References
# https://www.cdc.gov/malaria/about/faqs.html

# load("/Users/jchau/Documents/Learnings/STA 2102 Applied Stats 2/A3/eviMean.RData")
# load("/Users/jchau/Documents/Learnings/STA 2102 Applied Stats 2/A3/gambiares.RData")
# https://www.jstatsoft.org/article/view/v063i12

eUrl = "http://pbrown.ca/teaching/astwo/data/eviMean.RData"
eFile = basename(eUrl)
if (!file.exists(eFile)) download.file(eUrl, eFile)
load(eFile)
```

```
data("gambiaUTM")

gborder = raster::getData("GADM", country = "GMB", level = 0)
gborder = spTransform(gborder, projection(gambiaUTM))

myres = glgm(pos ~ phc + evi, data = gambiaUTM, grid = squareRaster(gborder,70),
             family = "binomial",
             prior = list(sd = c(u = 0.01,alpha = 0.05), range = c(65 * 1000, alpha=0.05)),
             covariates = list(evi = eviMean))

# myres = glgm(pos ~ phc + evi, data = gambiaUTM, grid = squareRaster(gborder, 100), family =

# gmap = openmap(gambiaUTM, fact = 2)
# gmap2 = tonerToTrans(openmap(gambiaUTM, path = "stamen-toner",
# fact = 2))
# map.new(myres$raster, legendRight = 0.9)
# plot(gmap, add = TRUE)
# plot(gambiaUTM, add = TRUE)
# scaleBar(gambiaUTM, "bottom", cex = 1.2, bty = "n")

# par(mfrow=c(1,2))
# matplot(myres$parameters$sd$posterior[,'x'], myres$parameters$sd$posterior[,'y'],
#   type="l", lty=1, xlab = 'sd', ylab='density')
# legend('topright', lty=1, col=1:2, legend=c('posterior','prior'))
# lines(myres$parameters$sd$posterior[,'x'],myres$parameters$sd$posterior[,'prior'], col='red'
#
# matplot(myres$parameters$range$postK[,'x'], myres$parameters$range$postK[,c('y','prior')],
#   type="l", lty=1, xlab = 'range', ylab='density')
# legend('topright', lty=1, col=1:2, legend=c('posterior','prior'))

# rownames(myres$parameters$summary) = c('Intercept', 'Public Health Centre', 'Vegetation Inde
# knitr::kable(myres$parameters$summary[, c(1, 3, 5, 8)], digits = 5, escape = FALSE, format =
#              caption = 'Summary of GLGM Results')


# fitCol = colourScale(myres$raster[["predict.invlogit"]],
# style = "equal", breaks = 10, dec = -log10(0.02), col = "RdYlGn",
# rev = TRUE, opacity = 0.8)
# map.new(myres$raster, legendRight = 0.9)
# plot(myres$raster[["predict.invlogit"]], add = TRUE, col = fitCol$colOpacity,
# breaks = fitCol$breaks, legend = FALSE, main= "Expectation of Malaria")
# plot(gmap2, add = TRUE, maxpixels = 10^6)
# legendBreaks("right", fitCol, outer = TRUE, bty = "n", inset = 0)
#
# fitCol = colourScale(myres$raster[["random.mean"]],
#   style = "equal", breaks = 10, dec = -log10(0.02), col = "Accent",
#   rev = TRUE, opacity = 0.8)
# map.new(myres$raster, legendRight = 0.9)
# plot(myres$raster[["random.mean"]], add = TRUE, col = fitCol$colOpacity,
#      breaks = fitCol$breaks, legend = FALSE)
# plot(gmap2, add = TRUE, maxpixels = 10^10)
# legendBreaks("right", fitCol, outer = TRUE, bty = "n", inset = 0)

#
# fitCol = colourScale(myres$raster[["predict.exp"]],
#   style = "equal", breaks = 10, dec = -log10(0.02), col = "RdYlGn",
#   rev = TRUE, opacity = 0.8)
# map.new(myres$raster, legendRight = 0.9)
# plot(myres$raster[["predict.exp"]], add = TRUE, col = fitCol$colOpacity,
#      breaks = fitCol$breaks, legend = FALSE)
# plot(gmap2, add = TRUE, maxpixels = 10^10)
# legendBreaks("right", fitCol, outer = TRUE, bty = "n", inset = 0)
```