

Sujet 1 : Concentration de CO2 dans l'atmosphère depuis 1958

LL

Préparation des données

Les données de la concentration de CO2 dans l'atmosphère à l'observatoire de Mauna Loa, Hawaii, États-Unis, initié par Charles David Keeling en 1958, sont disponibles depuis le site Web de l'institut Scripps. Les observations hebdomadaires sont récupérées sous la forme d'un fichier au format CSV. Le téléchargement a été effectué le 25 novembre 2024. L'URL est:

```
data_url =  
"https://scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/monthly/monthly_in_situ_co2"
```

Nous faisons une copie locale de ce jeu de données car ce fichier est régulièrement mis à jour, nous téléchargeons les données seulement si la copie locale n'existe pas.

```
data_file = "co2.csv"  
if (!file.exists(data_file)) {  
  download.file(data_url, data_file, method="auto")  
}
```

Voici l'explication des colonnes incluse dans le jeu de données :

Nom de colonne	Libellé de colonne
Yr, Mn, Date, Date.1	donnent la date dans des formats différents
C02	concentration mensuelle de CO2 à Mauna Loa en micro-mol de CO2 par mole (ppm), ajustée sur 24:00 heures pour le 15ème jour de chaque mois
seasonally	même donnée que la colonne C02 mais avec le cycle saisonnier retiré
fit	données lissées
seasonally.1	données lissées avec saisonnalité retirée
C02.1	même donnée que colonne C02 mais avec les valeurs manquantes remplacées par les données de la colonne fit
seasonnally.2	même données que seasonnally avec valeurs manquantes remplacées par valeurs de seasonnally.1 , les valeurs manquantes sont notées -99.99
Sta	identifiant en 3 lettres de la station d'échantillonnage, MLO pour l'observatoire de Mauna Loa et MKO pour le sommet du Maunakea dont les données ont été utilisées pour combler les trous laissés entre décembre 2022 et février 2023 par l'éruption de Mauna Loa.

Les 61 premières lignes du fichier CSV sont des commentaires sur le contenu du fichier, la manière de citer, etc... que nous ignorons en précisant `skip=61`.

Lecture

```
data = read.csv(data_file, skip=61)
```

Regardons ce que nous avons obtenu:

```
head(data)
```

```
##      Yr Mn      Date  Date.1      CO2 seasonally      fit seasonally.1
## 1   NA NA          NA          adjusted          adjusted fit
## 2   NA NA      Excel      NA      [ppm]      [ppm]      [ppm]      [ppm]
## 3 1958  1    21200 1958.041   -99.99   -99.99   -99.99   -99.99
## 4 1958  2    21231 1958.126   -99.99   -99.99   -99.99   -99.99
## 5 1958  3    21259 1958.203   315.71   314.43   316.20   314.91
## 6 1958  4    21290 1958.288   317.45   315.15   317.31   314.99
##      CO2.1      seasonally.2 Sta
## 1      filled adjusted filled
## 2      [ppm]      [ppm]
## 3      -99.99      -99.99 MLO
## 4      -99.99      -99.99 MLO
## 5      315.71      314.43 MLO
## 6      317.45      315.15 MLO
```

```
tail(data)
```

```
##      Yr Mn      Date  Date.1      CO2 seasonally      fit seasonally.1
## 801 2024  7    45488 2024.538   425.40   424.63   425.20   424.46
## 802 2024  8    45519 2024.623   422.70   424.30   423.13   424.76
## 803 2024  9    45550 2024.708   421.59   425.11   421.53   425.07
## 804 2024 10    45580 2024.790   422.05   425.66   421.77   425.37
## 805 2024 11    45611 2024.874   -99.99   -99.99   -99.99   -99.99
## 806 2024 12    45641 2024.956   -99.99   -99.99   -99.99   -99.99
##      CO2.1      seasonally.2 Sta
## 801      425.40      424.63 MLO
## 802      422.70      424.30 MLO
## 803      421.59      425.11 MLO
## 804      422.05      425.66 MLO
## 805      -99.99      -99.99 MLO
## 806      -99.99      -99.99 MLO
```

Modification des noms de colonnes

Les deux premières lignes correspondent à la suite des noms de colonnes qui ont été interprétées comme observations. Nous devons donc ramener les éléments de ces deux lignes dans les noms de colonnes. Pour cela nous créons un vecteur vide `new_colnames` qui permettra de stocker nos nouveaux noms de variables. Nous pouvons remarquer que certaines cases contiennent des `NA`, si c'est le cas il ne sert à rien de les coller au nom de colonne d'origine et nous passons ces cases. Certaines cases sont remplies d'espaces mais ne contiennent pas de `NA` donc nous passons également ces cases. Nous fonctionnons de manière itérative en collant les cases à chaque nom de colonne puis nous ajoutons le nouveau nom de la colonne au vecteur `new_colnames`. Nous vérifions les nouveaux noms de colonne en affichant le vecteur final.

```

new_colnames <- vector()
for (j in 1:length(colnames(data))){
  x <- colnames(data)[[j]]
  for (i in 1:2){
    if (is.na(data[i,j])){
      next
    }else if(data[i,j] == " "){
      next
    }else{
      x <- gsub("\\s+", " ", paste(x, data[i,j]))
    }
  }
  new_colnames <- c(new_colnames, x)
}
print(new_colnames)

```

```

## [1] "Yr" "Mn"
## [3] "Date Excel" "Date.1"
## [5] "CO2 [ppm]" "seasonally adjusted [ppm] "
## [7] "fit [ppm]" "seasonally.1 adjusted fit [ppm]"
## [9] "CO2.1 filled [ppm]" "seasonally.2 adjusted filled [ppm]"
## [11] "Sta "

```

Les nouveaux noms semblent corrects, nous allons à présent les indiquer comme nouveaux noms de colonnes puis nous supprimerons les deux premières lignes car elles ne sont plus utiles.

```

colnames(data) <- new_colnames
data <- data[-c(1:2),]

```

Nous pouvons vérifier l'apparence du début de notre jeu de données, les noms ont bien été modifié et les 2 premières lignes retirées.

```
head(data)
```

```

##      Yr Mn Date Excel  Date.1 CO2 [ppm] seasonally adjusted [ppm]  fit [ppm]
## 3 1958  1      21200 1958.041  -99.99                -99.99      -99.99
## 4 1958  2      21231 1958.126  -99.99                -99.99      -99.99
## 5 1958  3      21259 1958.203   315.71                314.43      316.20
## 6 1958  4      21290 1958.288   317.45                315.15      317.31
## 7 1958  5      21320 1958.370   317.51                314.69      317.89
## 8 1958  6      21351 1958.455  -99.99                -99.99      317.27
##      seasonally.1 adjusted fit [ppm] CO2.1 filled [ppm]
## 3                      -99.99                -99.99
## 4                      -99.99                -99.99
## 5                      314.91                315.71
## 6                      314.99                317.45
## 7                      315.07                317.51
## 8                      315.15                317.27
##      seasonally.2 adjusted filled [ppm] Sta
## 3                      -99.99  MLO
## 4                      -99.99  MLO
## 5                      314.43  MLO

```

```
## 6          315.15  MLO
## 7          314.69  MLO
## 8          315.15  MLO
```

Types des colonnes

L'observation des premières lignes de données mettent en évidence que les colonnes de concentration de CO2 (colonnes 5 à 10) sont toutes considérées comme de type `character` qui est probablement dû aux deux premières lignes qui ont été retirées et qui contenaient des mots. Nous devons donc changer le type de ces colonnes pour le type `numeric`.

```
data[,5:10] <- lapply(data[,5:10], as.numeric)
```

Valeurs manquantes

De plus, nous voyons dans les premières lignes de données des observations avec des valeurs de CO2 de -99.99. Les commentaires du fichier indiquent que les valeurs manquantes sont indiquées par cette valeur. Il faudrait donc convertir tous les -99.99 en NA afin de les traiter comme tels.

```
data[data == -99.99] <- NA
```

Maintenant, nous pouvons mesurer la quantité de NA dans nos données.

```
na_records = apply(data, 1, function (x) any(is.na(x)))
data[na_records,]
```

```
##      Yr Mn Date Excel   Date.1 C02 [ppm] seasonally adjusted [ppm] fit [ppm]
## 3   1958  1    21200 1958.041      NA              NA              NA
## 4   1958  2    21231 1958.126      NA              NA              NA
## 8   1958  6    21351 1958.455      NA              NA             317.27
## 12  1958 10    21473 1958.789      NA              NA             312.42
## 76  1964  2    23422 1964.126      NA              NA             320.04
## 77  1964  3    23451 1964.205      NA              NA             320.76
## 78  1964  4    23482 1964.290      NA              NA             321.84
## 805 2024 11    45611 2024.874      NA              NA              NA
## 806 2024 12    45641 2024.956      NA              NA              NA
##      seasonally.1 adjusted fit [ppm] C02.1 filled [ppm]
## 3              NA              NA
## 4              NA              NA
## 8              315.15           317.27
## 12             315.41           312.42
## 76             319.37           320.04
## 77             319.42           320.76
## 78             319.46           321.84
## 805            NA              NA
## 806            NA              NA
##      seasonally.2 adjusted filled [ppm] Sta
## 3              NA  MLO
## 4              NA  MLO
## 8             315.15  MLO
## 12            315.41  MLO
```

```
## 76          319.37 MLO
## 77          319.42 MLO
## 78          319.46 MLO
## 805         NA MLO
## 806         NA MLO
```

Nous avons ainsi 9 lignes contenant des **NA** avec deux cas précis : - soit toutes les cases des colonnes de valeurs numériques sont des **NA**, c'est le cas des observations de janvier et février 1958 (les mesures ont commencées en mars 1958) et de novembre et décembre 2024 (pas encore écoulés au moment de l'analyse) - soit les deux premières colonnes présentent effectivement des **NA** mais les autres contiennent des valeurs calculées à partir du lissage des données permettant de remplir les cases de ces observations à l'origine manquantes. (information donnée dans les commentaires du fichier)

Nous savons donc à présent que 4 mois (janvier et février 1958, novembre et décembre 2024) ne pourront pas être utilisés, pour les autres il sera possible d'utiliser une valeur approximée par lissage. Nous allons donc retirer les deux premières (janvier et février 1958) et les deux dernières (novembre et décembre 2024) lignes.

```
data <- data[-c(1,2,803,804),]
```

Conversion des dates

Les deux colonnes de dates qui nous intéressent sont **Yr** pour l'année et **Mn** pour le mois. Nous pouvons vérifier leurs classes:

```
class(data$Yr)
```

```
## [1] "integer"
```

```
class(data$Mn)
```

```
## [1] "integer"
```

Puisque nous travaillons sur une série temporelle, il est possible de créer un objet de type **ts** pour "time series" avec lequel il sera simple d'effectuer des analyses. Pour cela, il suffit d'utiliser la fonction **ts** avec la colonne 9, qui contient les valeurs de la colonne 5 mais avec les valeurs manquantes repliés par lissage. L'argument **frequency** prend la valeur 12 car c'est une série mensuelle, l'argument **start** commence à la première ligne du tableau avec l'année contenue dans la colonne **Yr** et le mois dans la colonne **Mn**.

```
TS_CO2 <- ts(data = data$`C02.1 filled [ppm]`, frequency = 12, start = c(data[1, "Yr"],
  data[1, "Mn"]))
class(TS_CO2)
```

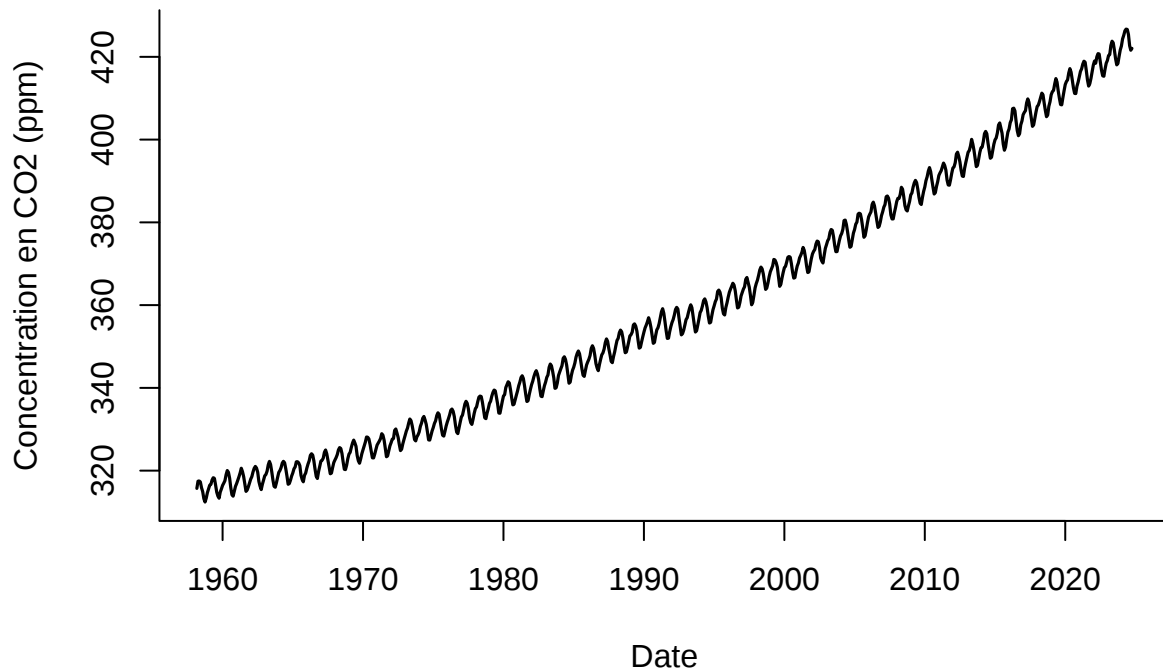
```
## [1] "ts"
```

Inspection

Nous pouvons à présent regarder à quoi ressemblent nos données en visualisant notre objet de type série temporelle, c'est à dire la concentration de CO2 en fonction du temps.

```
plot(TS_CO2, type="l", xlab="Date", ylab="Concentration en CO2 (ppm)", main = "Evolution de la concentr
```

Evolution de la concentration en CO2 (ppm) à Mauna Loa depuis 1958 (valeurs manquantes remplies)

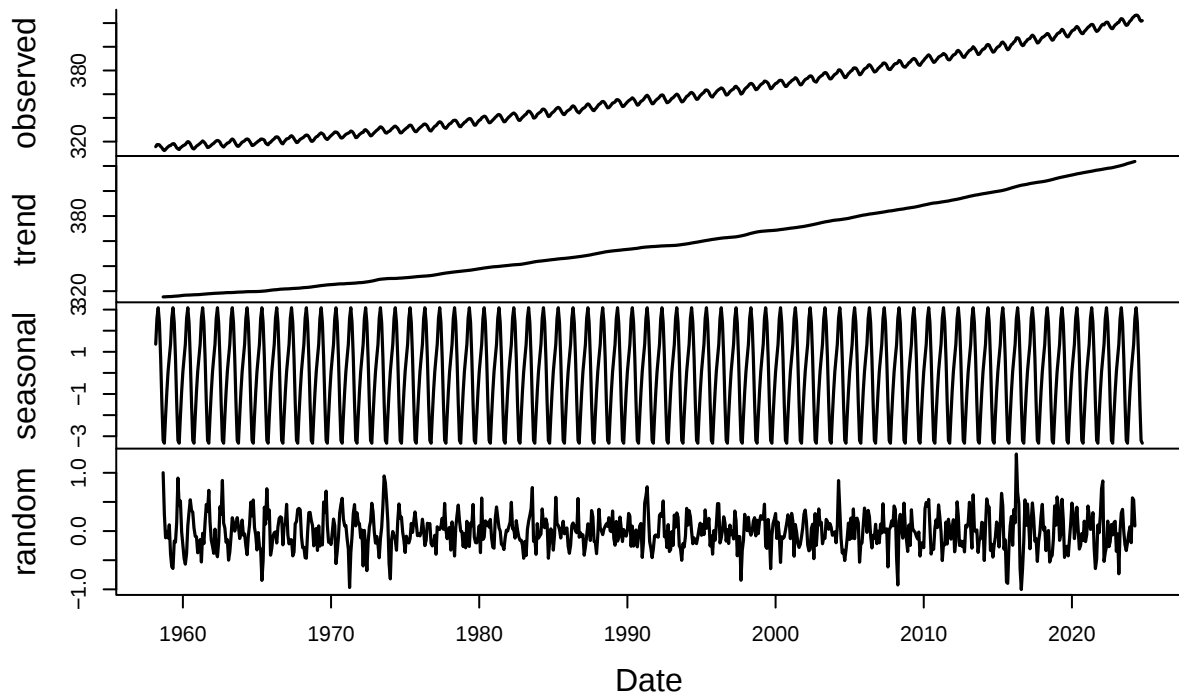


Nous retrouvons la fameuse courbe mettant en évidence l'évolution systématique et l'oscillation périodique de la concentration en CO2 à Mauna Loa depuis 1958.

La fonction `decompose` permet de décomposer une série temporelle en séparant la tendance, la saisonnalité/périodicité et les éléments de variation aléatoire. L'objet décomposé peut ensuite être visualisé.

```
TS_CO2_decomp <- decompose(TS_CO2)
plot(TS_CO2_decomp, xlab="Date", lwd = 1.6, bty = "l")
```

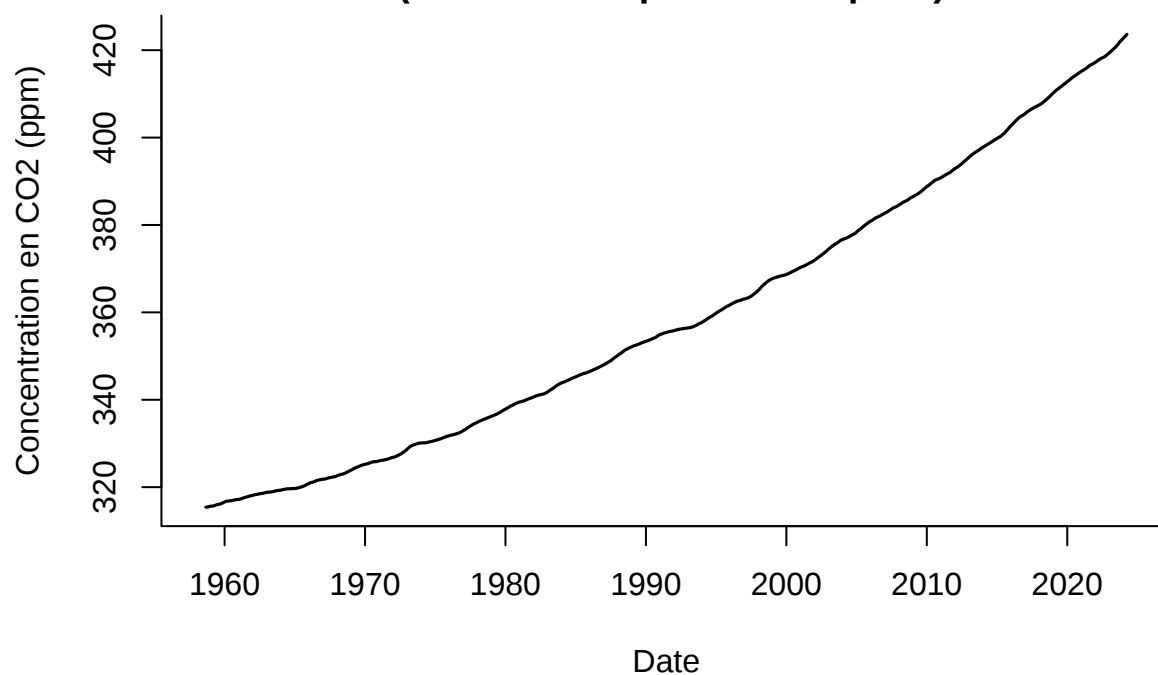
Decomposition of additive time series



Nous pouvons visualiser la tendance et la périodicité séparément :

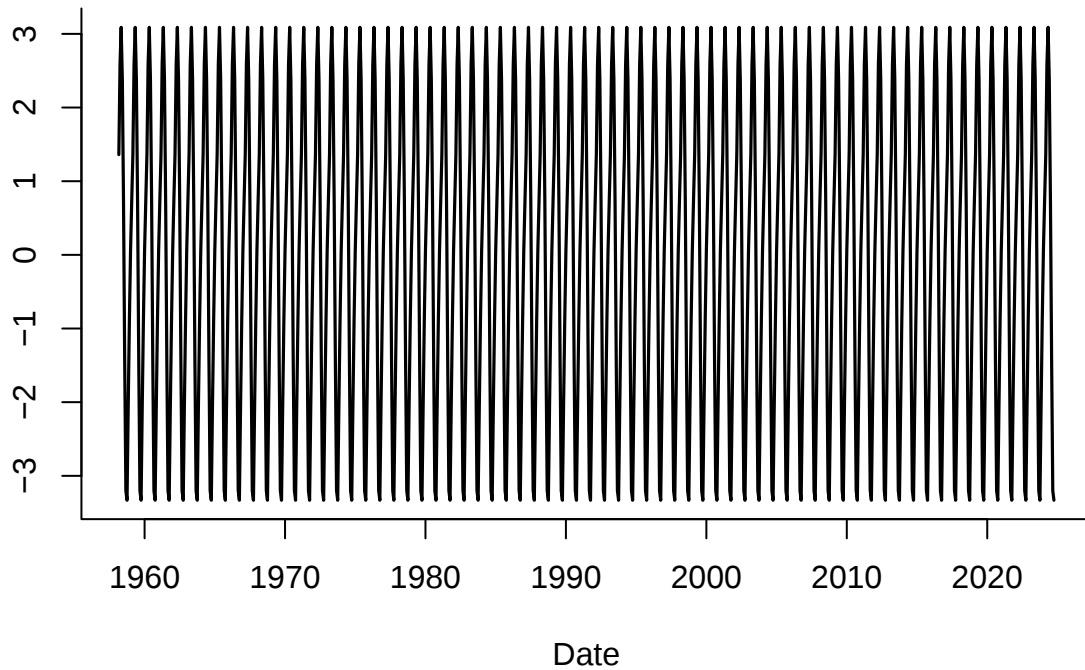
```
plot(TS_CO2_decomp$trend, xlab="Date", ylab="Concentration en CO2 (ppm)", main = "Tendance de l'évolution")
```

**Tendance de l'évolution de la concentration en CO2 (ppm)
à Mauna Loa depuis 1958
(valeurs manquantes remplies)**



```
plot(TS_CO2_decomp$seasonal, xlab="Date", ylab = "", main = "Périodicité de la concentration en CO2 (ppm)",
```

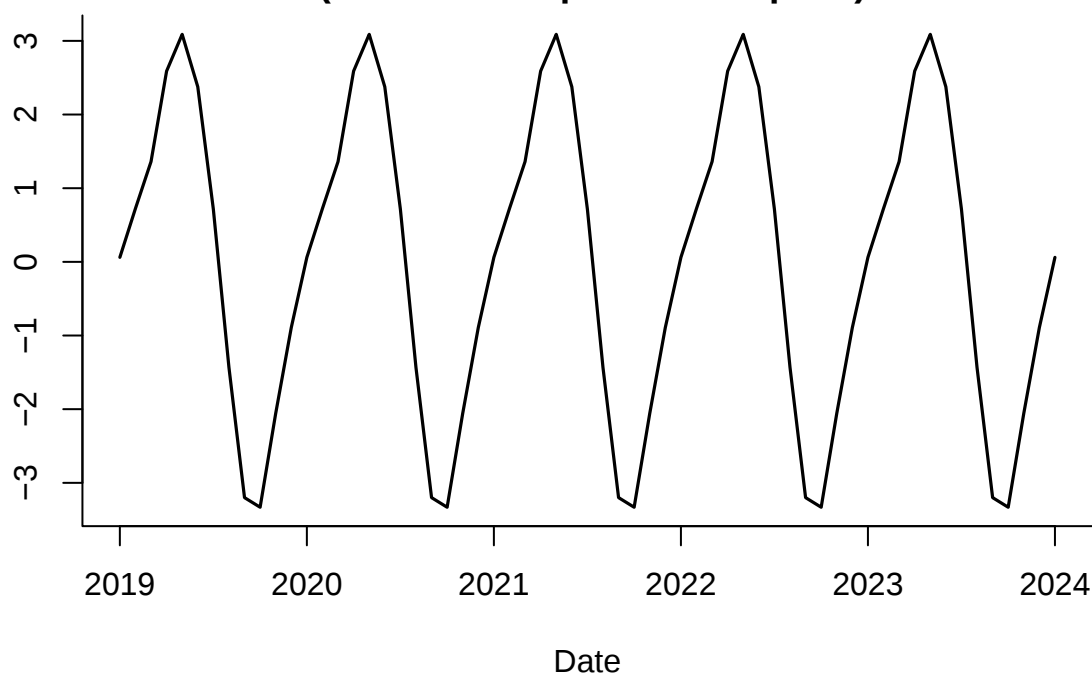
**Périodicité de la concentration en CO2 (ppm)
à Mauna Loa depuis 1958
(valeurs manquantes remplies)**



Nous pouvons sélectionner un sous-ensemble des 5 dernières années (complètes) pour mieux visualiser la périodicité de la concentration en CO2.

```
plot(window(TS_CO2_decomp$seasonal, start = 2019, end = 2024), xlab="Date", ylab="", main = "Périodicité")
```

**Périodicité de la concentration en CO₂ (ppm)
à Mauna Loa de 2019 à 2024
(valeurs manquantes remplies)**



Nous pouvons observer que les pics semblent se produire à la fin de l'hiver/ au printemps et les creux à la fin de l'été/à l'automne.