Analyse de l'effet du changement climatique

Étude descriptive et exploratoire

Manar Trimeche

6 janvier 2026

Table des matières

1 Introduction

Le changement climatique est un enjeu majeur de santé publique. Ce mini-projet a pour objectif d'étudier l'effet d'événements climatiques (vagues de chaleur, inondations, vagues de froid, contamination de l'eau, etc.) sur la santé d'une population urbaine et rurale à partir du jeu de données `Effect_Climate_Change.csv`.

Plus précisément, l'analyse vise à :

— Décrire les caractéristiques socio-démographiques, climatiques et sanitaires de la population.
— Explorer les liens entre événements climatiques et présence de maladies.
— Estimer des modèles simples pour identifier les principaux facteurs associés à la maladie.

2 Préparation de l'environnement

Les analyses statistiques ont été réalisées avec le langage de programmation **R** et les packages suivants :

```
# Installation des packages n cessaires
install.packages(c("tidyverse", "ggplot2", "dplyr",
                   "corrplot", "psych", "car"))

# Chargement des librairies
library(tidyverse)
library(ggplot2)
library(dplyr)
library(corrplot)
library(psych)
library(car)
```

— **tidyverse** : manipulation et visualisation de données
— **ggplot2** : graphiques avancés
— **dplyr** : manipulation de données
— **corrplot** : visualisation de corrélations
— **psych** : statistiques descriptives
— **car** : tests statistiques (VIF, ANOVA)

3 Chargement et exploration des données

3.1 Aperçu des données

Le jeu de données `Effect_Climate_Change.csv` contient :

— **800 observations** (individus)
— **31 variables** incluant :
  — Variables démographiques : âge, sexe, région, ville
  — Variables économiques : revenu mensuel
  — Variables de santé : BMI, catégorie de BMI, présence de maladies, degré de maladie
  — Variables climatiques : vagues de chaleur, inondations, vagues de froid, contamination de l'eau, hivers rudes

```
# Charger le dataset
data <- read.csv("Effect_Climate_Change.csv",
                 header = TRUE,
                 stringsAsFactors = TRUE)

# Aper u
```

```r
head(data)
dim(data)       # 800 x 31
str(data)       # Structure
summary(data)   # Statistiques descriptives
```

4 Nettoyage des données

Les étapes de nettoyage ont été :

1. **Suppression du SNO** : colonne de numéro de série inutile
2. **Normalisation des noms** : application de `make.names()` pour compatibilité R
3. **Conversion en facteurs** : 13 variables catégorielles transformées

```r
# Nettoyage
data_clean <- data %>% select(-SNO)
names(data_clean) <- make.names(names(data_clean))

# Conversion des cat goriques
categorical_vars <- c("REGION.", "SEX", "MARITAL.STATUS",
                      "EDUCATION", "OCCUPATION", "BMI.CATEGORY",
                      "DISEASE", "DISEASE.TYPE", "FLOODS",
                      "CONTAMINATION.OF.WATER", "COLD.WAVES",
                      "WINTER.STORMS", "CITY")

for (var in categorical_vars) {
    if (var %in% names(data_clean)) {
        data_clean[[var]] <- as.factor(data_clean[[var]])
    }
}
```

5 Statistiques descriptives

5.1 Variables numériques

Les statistiques descriptives des variables numériques clés sont :

TABLE 1 – Statistiques descriptives des variables numériques

Variable	Min	Q1	Médiane	Moyenne	Q3	Max
AGE	18	35	50	48.5	62	80
WEIGHT (kg)	45	65	75	73.2	82	130
HEIGHT (cm)	140	163	170	169.8	177	200
BMI	15	23	26	25.3	28	42
HEAT.WAVES	0	5	8	7.8	11	20
DEGREE.OF.DISEASE	0	2	4	4.2	6	10

5.2 Variables catégorielles

Distribution par région, sexe, maladie, inondations et ville :

TABLE 2 – Distribution des variables catégorielles principales

Variable	Catégorie	Fréquence (%)
2*SEXE	Homme	45%
	Femme	55%
2*MALADIE	Oui	38%
	Non	62%
2*INONDATIONS	Oui	52%
	Non	48%

6 Analyses de corrélation

6.1 Matrice de corrélation

Une matrice de corrélation de Pearson a été calculée entre les variables numériques clés.

6.1.1 Formule

La corrélation entre deux variables X et Y est :

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

6.2 Corrélation vagues de chaleur – degré de maladie

```
cor.test(data_clean$HEAT.WAVES, data_clean$DEGREE.OF.DISEASE)
```

Ce test fournit :
— Le coefficient de corrélation r de Pearson
— La statistique de test t
— La valeur-p (significativité)
— L'intervalle de confiance à 95%

7 Tests statistiques

7.1 Test du Chi-carré

7.1.1 Inondations vs maladies

Hypothèses :
— H_0 : Les inondations et les maladies sont indépendantes
— H_1 : Les inondations et les maladies sont associées
Statistique de test :

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où O_{ij} sont les fréquences observées et E_{ij} les fréquences attendues.

```
table_floods_disease <- table(data_clean$FLOODS,
                              data_clean$DISEASE)
chi_test_1 <- chisq.test(table_floods_disease)
chi_test_1
```

7.1.2 Contamination de l'eau vs maladies

Un test similaire est appliqué pour la contamination de l'eau.

7.2 Test t de Student

7.2.1 BMI selon exposition aux inondations

Hypothèses :
— H_0 : $\mu_{\text{Inondations=Oui}} = \mu_{\text{Inondations=Non}}$
— H_1 : Les moyennes de BMI diffèrent selon l'exposition aux inondations
Statistique de test :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

```
t_test_floods <- t.test(BMI ~ FLOODS, data = data_clean)
```

7.3 ANOVA

7.3.1 Vagues de chaleur par ville

$$F = \frac{\text{Variance inter-groupes}}{\text{Variance intra-groupes}}$$

```
anova_city <- aov(HEAT.WAVES ~ CITY, data = data_clean)
summary(anova_city)
tukey_test <- TukeyHSD(anova_city)
```

Les comparaisons multiples de Tukey permettent d'identifier quelles villes diffèrent significativement.

7.3.2 Degré de maladie par catégorie BMI

```
data_disease_only <- data_clean %>% filter(DISEASE == "Yes")
anova_bmi <- aov(DEGREE.OF.DISEASE ~ BMI.CATEGORY,
                 data = data_disease_only)
summary(anova_bmi)
```

8 Modèles de régression

8.1 Régression linéaire simple

8.1.1 Modèle

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

où :
— Y_i : degré de maladie (parmi les malades)
— X_i : intensité des vagues de chaleur
— ϵ_i : résidu

```
model_1 <- lm(DEGREE.OF.DISEASE ~ HEAT.WAVES,
              data = data_disease_only)
summary(model_1)
```

8.1.2 Interprétation

Les résultats fournissent :
— Le coefficient de l'ordonnée à l'origine $\hat{\beta}_0$
— Le coefficient de la pente $\hat{\beta}_1$ (effet de HEAT.WAVES)
— L'erreur-type de chaque coefficient
— La statistique t et valeur-p
— R^2 : proportion de variance expliquée

8.2 Régression linéaire multiple

8.2.1 Modèle

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

avec :
— X_1 : vagues de chaleur
— X_2 : BMI
— X_3 : âge

```
model_2 <- lm(DEGREE.OF.DISEASE ~ HEAT.WAVES + BMI + AGE,
              data = data_disease_only)
summary(model_2)

# V rification de la multicolin arit
vif(model_2)
```

8.2.2 Diagnostics

— **VIF (Variance Inflation Factor)** : détecte la multicolinéarité
— **Graphiques de résidus** : vérification des hypothèses (normalité, homoscédasticité)
— R^2 **ajusté** : comparaison avec le modèle simple

8.3 Régression logistique

8.3.1 Modèle

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

ou en termes de probabilité :

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

```
# Cr er une variable binaire pour la maladie
data_clean$DISEASE_binary <- ifelse(data_clean$DISEASE == "Yes", 1, 0)

# Cr er des versions discr tes pour la r gression logistique
data_clean$FLOODS.descrete <- ifelse(data_clean$FLOODS == "Yes", 1, 0)
data_clean$CONTAMINATION.OF.WATER.descrete <- ifelse(data_clean$
    CONTAMINATION.OF.WATER == "Yes", 1, 0)

# Mod le logistique
model_logistic <- glm(DISEASE_binary ~ HEAT.WAVES + BMI + AGE +
```

```
                          FLOODS.descrete + CONTAMINATION.OF.WATER.
                              descrete,
                        data = data_clean,
                        family = binomial)
summary(model_logistic)

# Odds ratios
exp(coef(model_logistic))
```

8.3.2 Interprétation des Odds Ratios

L'**Odds Ratio** (OR) représente le rapport des cotes :

$$OR = e^{\beta}$$

— $OR = 1$: pas d'effet
— $OR > 1$: augmentation du risque de maladie
— $OR < 1$: diminution du risque de maladie

Par exemple, si $OR = 1.05$ pour HEAT.WAVES, une augmentation d'une unité des vagues de chaleur multiplie les odds de maladie par 1.05 (augmentation de 5%).

9 Conclusion

Ce rapport a présenté une analyse descriptive et exploratoire de l'effet du changement climatique sur la santé.

9.1 Résultats clés

— Les statistiques descriptives révèlent l'âge moyen, le BMI moyen et la proportion de malades dans la population.
— Les analyses de corrélation identifient les liens entre variables climatiques et sanitaires.
— Les tests statistiques évaluent la significativité des associations.
— Les modèles de régression quantifient l'impact des facteurs climatiques sur la santé.

9.2 Limitations

— Étude transversale (pas de causalité établie)
— Contexte spécifique (population urbaine/rurale)
— Variables observationnelles (pas de contrôle expérimental)
— Potentielles variables confondantes non mesurées

9.3 Perspectives futures

— Études longitudinales pour évaluer la causalité
— Intégration de variables supplémentaires (pollution, comportements)
— Modèles plus complexes (interaction, non-linéarité)
— Analyse stratifiée par groupes de population

A Ressources

— R Core Team (2023). *R : A Language and Environment for Statistical Computing*
— Wickham, H. (2016). *ggplot2 : Elegant Graphics for Data Analysis*
— James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*

B Code R complet

Le code R complet se trouve dans le fichier Quarto (`.qmd`) accompagnant ce rapport.