



**MINISTÈRE DE
L'ENSEIGNEMENT SUPÉRIEUR,
DE LA RECHERCHE ET DE
L'INNOVATION**

**UNIVERSITÉ JOSEPH KI-
ZERBO**

**INSTITUT SUPERIEUR DES
SCIENCES DE LA POPULATION**

LICENCE PROFESSIONNELLE EN ANALYSE STATISTIQUE

PROJET QUANTITATIF :modélisation hédonique

Membres du groupe :

NIKIEMA Tahirou

SANKARA Saïdou

TONGO Lazare

ENSEIGNANT :

Dr Fabrice YAMEOGO

I. Table des matières

INTRODUCTION

II.	Contexte et Justification.....	3
III.	Étapes préliminaires.....	3
IV.	Analyse Descriptive des données	5
1.	summary.....	5
2.	Variable de Distributions	5
3.	correlation	6
4.	selection des variables explicatives.....	7
5.	Transformation de variables	7
V.	Regression par modèle Lineaire multiple.....	9
1.	Sélectionner les variables explicatives	9
2.	selection du meilleur modèle summary(modele_selectionne)	9
3.	Test d'hypothèses.....	10
a.	Autocorrelation	10
b.	Homoscedasticité	11
c.	Normalité des residus.....	11
d.	Colinearité	11
e.	Conclusion sur les hypothèses	12
VI.	Régression Lasso.....	12
1.	Configuration des données	13
2.	Les graphiques des coefficients.....	14
3.	Ajuster le modèle LASSO sur l'ensemble des données	15
4.	Interprétation des Estimations du Logarithme du Coût des Parcelles.....	17
5.	Prédire les valeurs de InCOUT pour toutes les années	18
6.	Les indices élémentaires ou simples	19
a.	Interpretation des indices des prix.....	20
b.	Interpretraion des indices de valeurs.....	23
c.	Interpretation des indices de Laspeyres	24
d.	Interpretation des indices de Paasche	25
e.	Interpretation des indices de Fischer	26

CONCLUSION

Modelisation Hedonique

II. Contexte et Justification

La valeur des terrains à Ouagadougou a considérablement évolué entre 2018 et 2024, en raison de l'urbanisation et des améliorations infrastructurelles. Le modèle hédonique décompose les prix des parcelles selon leurs caractéristiques, aidant ainsi à comprendre l'impact des politiques urbaines sur le marché foncier.

III. Étapes préliminaires

Assurez-vous que nous avons les packages nécessaires installés/chargés.

```
library(tidyverse)

## Warning in Sys.timezone(): unable to identify current timezone 'T':
## please set environment variable 'TZ'

## — Attaching core tidyverse packages ——————— tidyverse
2.0.0 —
## ✓ dplyr      1.1.4      ✓ readr      2.1.4
## ✓forcats    1.0.0      ✓ stringr    1.5.0
## ✓ ggplot2    3.5.1      ✓ tibble     3.2.1
## ✓ lubridate  1.9.2      ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## — Conflicts ——————— tidyverse_conflic
cts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()   masks stats::lag()
## [i] Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors

library(glmnet)

## Le chargement a nécessité le package : Matrix
##
## Attachement du package : 'Matrix'
##
## Les objets suivants sont masqués depuis 'package:tidyverse':
##
##       expand, pack, unpack
##
## Loaded glmnet 4.1-7

library(gridExtra)

##
## Attachement du package : 'gridExtra'
##
## L'objet suivant est masqué depuis 'package:dplyr':
```

```

##      combine

library(tidyr)
library(lubridate)
library(dplyr)
library(questionr)
library(car)

## Le chargement a nécessité le package : carData
##
## Attachement du package : 'car'
##
## L'objet suivant est masqué depuis 'package:dplyr':
##
##      recode
##
## L'objet suivant est masqué depuis 'package:purrr':
##
##      some

library(performance)
library(lmtest)

## Le chargement a nécessité le package : zoo
##
## Attachement du package : 'zoo'
##
## Les objets suivants sont masqués depuis 'package:base':
##
##      as.Date, as.Date.numeric

library(caret)

## Le chargement a nécessité le package : lattice
##
## Attachement du package : 'caret'
##
## L'objet suivant est masqué depuis 'package:purrr':
##
##      lift

```

Load data

```

setwd("D:/projet_quanti")
Parcelles <- read_delim("Parcelles.csv", delim = ";", escape_double = FALSE
, col_types = cols(Date_vente = col_date(format = "%d/%m/%Y"), Date_fin_contrat = col_date(format = "%d/%m/%Y")), trim_ws = TRUE)

```

Les données sont chargées depuis un fichier CSV et contiennent des informations sur 1811 parcelles, telles que la ville, le site, l'usage, la superficie, le coût par mètre carré, le coût total, les taxes, le type d'option de paiement, et diverses attestations et plans établis. Une exploration descriptive des données sera effectuée pour comprendre la distribution et les relations entre les variables.

IV. Analyse Descriptive des données

1. summary

```
summary(Parcelles)

##      Ville             Site            Usage           Superficie
##  Length:1811      Length:1811      Length:1811      Min.   : 82
##  Class :character  Class :character  Class :character  1st Qu.: 300
##  Mode  :character  Mode  :character  Mode  :character  Median  : 320
##                                         Mean   : 495
##                                         3rd Qu.: 488
##                                         Max.   :13611
##      Cout_m2          COUT          Taxe_Jouissance Type_option
##  Min.   : 0     Min.   :       0     Min.   : 0.0  Length:1811
##  1st Qu.:26000  1st Qu.: 7644000  1st Qu.:250.0  Class :character
##  Median :26000  Median : 8320000  Median :500.0  Mode  :character
##  Mean   :27350   Mean   :16068350  Mean   :514.8
##  3rd Qu.:36000  3rd Qu.:17107500  3rd Qu.:500.0
##  Max.   :190000  Max.   :722376000 Max.   :3000.0
##      Date_vente        Date_fin_contrat  attestation_etablie
##  Min.   :2018-01-18  Min.   :2019-02-14  Length:1811
##  1st Qu.:2020-11-03  1st Qu.:2022-07-31  Class :character
##  Median :2022-10-28  Median :2023-12-04  Mode  :character
##  Mean   :2022-04-25  Mean   :2023-08-26
##  3rd Qu.:2023-11-08 3rd Qu.:2024-11-08
##  Max.   :2024-03-21  Max.   :2025-04-23
##      plan_etablie    Presence_ONEA    Presence SONABEL
##  Length:1811      Length:1811      Length:1811
##  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character
##
##
```

2. Variable de Distributions

```
table(Parcelles$Usage,Parcelles$Site,exclude=NULL)

##
##                                     BASSINKO SITE - BA CISSIN 2020 - SITE G
##  COMMERCE                           0          9
##  COMMERCE A L'ANGLE                 63          0
##  COMMERCE ANGLE                     0          0
##  COMMERCE ANGLE 1 BITUME            34          0
##  COMMERCE ANGLE 2 VOIES             0          0
##  COMMERCE ORDINAIRE ANGLE          0          0
##  COMMUNAUTAIRE                     0          0
##  HABITATION                         0          0
##  HABITATION ANGLE                  0          0
##  STATION SERVICE                   0          0
##
##                                     OUAGA 2000 - SITE A OUAGA 2000 - SITE AA
##  COMMERCE                           21         37
##  COMMERCE A L'ANGLE                 0          98
```

```

## COMMERCE ANGLE 2
## COMMERCE ANGLE 1 BITUME 0
## COMMERCE ANGLE 2 VOIES 0
## COMMERCE ORDINAIRE ANGLE 5
## COMMUNAUTAIRE 4
## HABITATION 105
## HABITATION ANGLE 21
## STATION SERVICE 0
##
## SECTEUR 16 OUAGA SILMIOUGOU
## COMMERCE 0 37
## COMMERCE A L'ANGLE 0 32
## COMMERCE ANGLE 0 0
## COMMERCE ANGLE 1 BITUME 0 0
## COMMERCE ANGLE 2 VOIES 0 0
## COMMERCE ORDINAIRE ANGLE 0 13
## COMMUNAUTAIRE 0 0
## HABITATION 1 895
## HABITATION ANGLE 0 193
## STATION SERVICE 0 0

Date_vente))

##
## 2018 2019 2020 2021 2022 2023 2024
## 168 104 219 78 349 823 70

```

3. correlation

```

cor(Parcelles$Cout_m2,Parcelles$COUT)

## [1] 0.5981968

cor(Parcelles$COUT,Parcelles$Superficie)

## [1] 0.6848416

cor(Parcelles$COUT,Parcelles$Taxe_Jouissance)

## [1] 0.4817309

```

Le choix des caractéristiques et l'endogénéité

La littérature sur les prix hédoniques ne définit que peu de règles explicites, impliquant généralement de travailler avec les données disponibles sans chercher à expliquer les prix implicites obtenus. La théorie économique conseille de ne pas inclure de variables

d'offre et de demande, car la fonction de prix hédonique représente l'interaction entre vendeurs et acheteurs, sans refléter directement ces variables. Des signes contraires à l'intuition peuvent apparaître en raison des relations complexes entre cette fonction et les courbes d'offre et de demande. Pakes (2003) relie cette fonction aux coûts marginaux des vendeurs, argument renforcé par le fait que l'endogénéité n'est pas un problème en prévision pure. Cependant, des incohérences peuvent résulter d'autres sources, comme l'omission de caractéristiques importantes ou l'utilisation de variables annexes.

4. selection des variables explicatives

L'analyse descriptive révèle que les variables 'présence ONEA', 'absence ONEA' et 'ville' n'ont qu'une seule modalité, ce qui signifie qu'elles n'influencent pas la variation du prix. De plus, la valeur minimale des coûts des terrains est de 0, ce qui n'est pas logique. Par conséquent, nous écartons les observations dont le coût du terrain est de 0. La théorie économique stipule que la fonction de prix hédonique ne doit pas inclure des variables qui décrivent directement l'offre et la demande ce qui nous amène à considérer que la date de fin de contrat et le type d'option n'influencent pas le prix de vente du terrain. Nous écartons donc ces variables de notre modélisation.

5. Transformation de variables

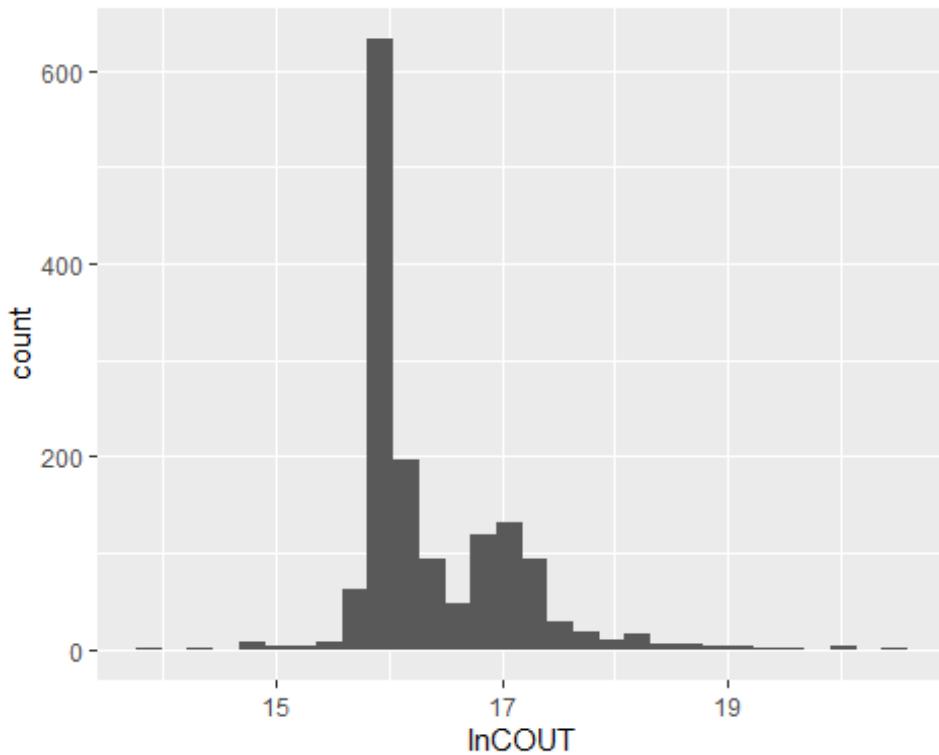
```
hdnc = Parcels %>%
  select(-c(Ville, Type_option, Date_fin_contrat, Presence_ONEA, Presence_S
ONABEL))%>%
  filter(COUT >0, Cout_m2 > 0) %>%
  mutate(years =as.character(year(Date_vente)),
        lnCOUT = log(COUT)
  )
```

Recodage de hdnc\$Site en hdnc\$Site_rec

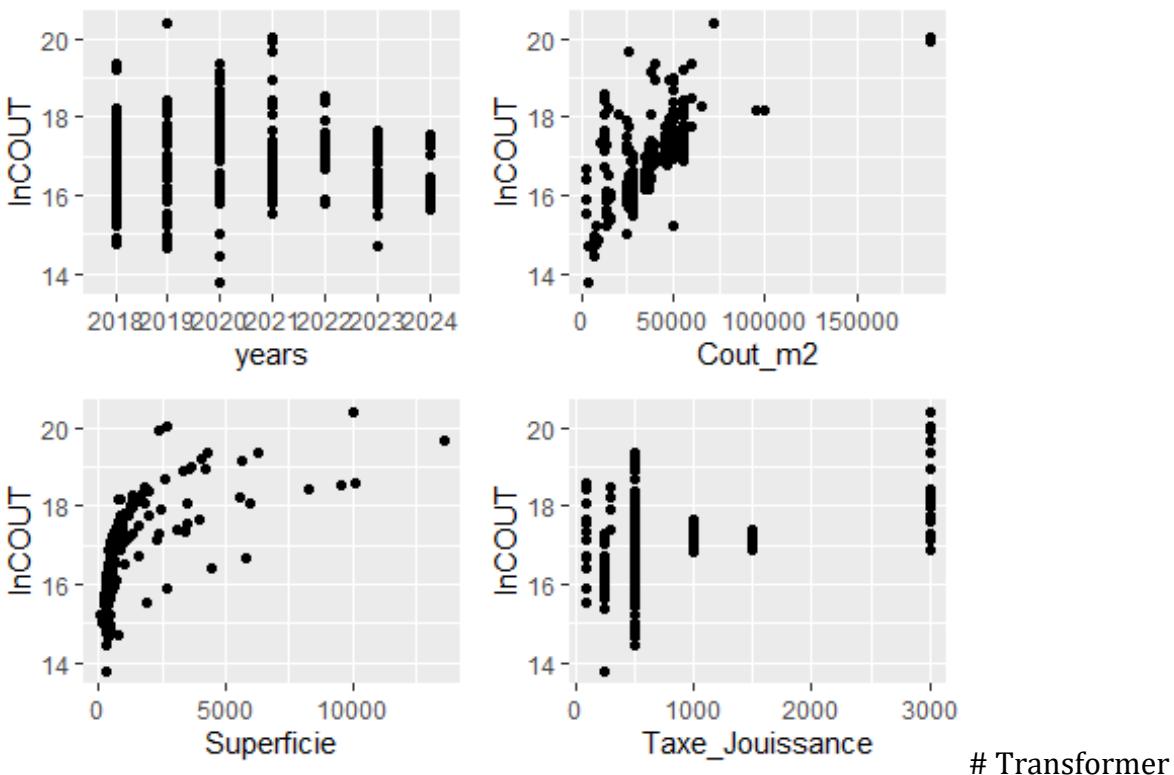
```
hdnc$Site_rec <- hdnc$Site
hdnc$Site_rec[hdnc$Site == "BASSINKO SITE - BA"] <- "BASSINKO_SITE_BA"
hdnc$Site_rec[hdnc$Site == "CISSIN 2020 - SITE G"] <- "CISSIN_2020_SITE_G"
hdnc$Site_rec[hdnc$Site == "OUAGA 2000 - SITE A"] <- "OUAGA_2000_SITE_A"
hdnc$Site_rec[hdnc$Site == "OUAGA 2000 - SITE AA"] <- "OUAGA_2000_SITE_AA"
hdnc$Site_rec[hdnc$Site == "SECTEUR 16 OUAGA"] <- "SECTEUR_16_OUAGA"
hdnc$Site_rec[hdnc$Site == "SILMIOUGOU"] <- "SILMIOUGOU"
```

Examinons la distribution des prix en logarithme et comment ils se rapportent (graphiquement) à certaines caractéristiques des maisons

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



et comment ils se rapportent (graphiquement) à certaines caractéristiques des parcelles



les variables catégorielles en variables de type facteur

```
hdnc = hdnc %>%
  mutate(Site_rec = as.factor(Site_rec), Usage= as.factor(Usage), plan_etable
lie = as.factor(plan_etable), attestation_etable = as.factor(attestation_
etable), years = as.factor(years))
```

Après le traitement des données, nous allons maintenant commencer l'implémentation de notre modèle hédonique. Cette implémentation débutera par l'ajustement d'un modèle de régression. À cet effet, nous allons essayer plusieurs modèles afin de déterminer celui qui s'adapte le mieux à nos données.

V. Regression par modèle Lineaire multiple

La première méthode est le Modèle Log-Linéaire utiliser pour calculer les indices hédoniques consiste à regrouper plusieurs périodes et à estimer un unique modèle hédonique auquel on ajoute des indicatrices temporelles. Lorsque l'ensemble des périodes sont prises en compte, le modèle est $\text{Log}(\text{Prix}) = \beta \cdot X + \delta \cdot D + \varepsilon$ avec X les caractéristiques et D les indicatrices temporelles (avec 2018 comme période de référence). Ce modèle est souvent utilisé pour traiter la distribution non normale des prix ce qui est conforme à nos données

1. Sélectionner les variables explicatives

Après revision de la litterature, les variables selectionnés sont : Usage, Superficie, Cout_m2, Taxe_Jouissance, attestation_etablie, plan_etablie, years, Site_rec En effet se sont les seuls variables que nous avons jugés representatives des caractéristiques physiques et des caractéristiques de quartier

NB : il n'y a pas de Caractéristiques environnementales dans nos données

```
modlm = lm(lnCOUT ~ Usage + Superficie + Cout_m2 + Taxe_Jouissance + attestation_etablie + plan_etablie + years + Site_rec, data = hdnc)
modele_selectionne <- step(modlm, direction = "backward")

## Start: AIC=-4374.41
## lnCOUT ~ Usage + Superficie + Cout_m2 + Taxe_Jouissance + attestation_etablie +
##       plan_etablie + years + Site_rec
##
##                                Df  Sum of Sq    RSS      AIC
## <none>                            79.783 -4374.4
## - attestation_etablie   2     0.654  80.438 -4366.1
## - plan_etablie          2     1.960  81.743 -4341.8
## - Taxe_Jouissance       1     2.348  82.132 -4332.7
## - years                 6     4.783  84.566 -4298.7
## - Usage                 9    13.209  92.992 -4161.5
## - Site_rec              4    21.928 101.711 -4016.5
## - Superficie            1    60.040 139.823 -3530.9
## - Cout_m2               1    87.238 167.021 -3263.0
```

2. selection du meilleur modèle

`summary(modele_selectionne)`

```
##
## Call:
## lm(formula = lnCOUT ~ Usage + Superficie + Cout_m2 + Taxe_Jouissance +
##      attestation_etablie + plan_etablie + years + Site_rec, data = hdnc)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -1.72751 -0.09411 -0.05478  0.12007  1.02281
```

```

## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.495e+01 7.884e-02 189.676 < 2e-16 ***
## UsageCOMMERCE A L'ANGLE -4.358e-02 3.456e-02 -1.261 0.207541
## UsageCOMMERCE ANGLE    -7.859e-01 1.767e-01 -4.447 9.38e-06 ***
## UsageCOMMERCE ANGLE 1 BITUME -1.738e-01 4.860e-02 -3.577 0.000358 ***
## UsageCOMMERCE ANGLE 2 VOIES   4.238e-01 7.063e-02  6.000 2.47e-09 ***
## UsageCOMMERCE ORDINAIRE ANGLE 9.976e-02 5.095e-02  1.958 0.050427 .
## UsageCOMMUNAUTAIRE      3.809e-01 8.175e-02  4.660 3.45e-06 ***
## UsageHABITATION        -2.055e-01 3.316e-02 -6.198 7.38e-10 ***
## UsageHABITATION ANGLE   -2.038e-01 3.567e-02 -5.712 1.35e-08 ***
## UsageSTATION SERVICE     2.461e-01 1.687e-01  1.459 0.144802
## Superficie              3.623e-04 1.086e-05 33.373 < 2e-16 ***
## Cout_m2                  3.827e-05 9.512e-07 40.228 < 2e-16 ***
## Taxe_Jouissance          1.470e-04 2.227e-05  6.600 5.70e-11 ***
## attestation_etableNON DEFINI -1.128e-01 4.168e-02 -2.708 0.006855 **
## attestation_etableOUI      5.156e-02 3.465e-02  1.488 0.136992
## plan_etableNON DEFINI      2.280e-01 3.798e-02  6.004 2.41e-09 ***
## plan_etableOUI             2.181e-02 2.767e-02  0.788 0.430648
## years2019                 -2.353e-02 2.986e-02 -0.788 0.430703
## years2020                 -1.843e-01 2.648e-02 -6.958 5.18e-12 ***
## years2021                 -2.743e-01 3.489e-02 -7.863 7.15e-15 ***
## years2022                 -1.796e-01 4.334e-02 -4.145 3.60e-05 ***
## years2023                 -3.503e-01 6.333e-02 -5.532 3.75e-08 ***
## years2024                 -3.598e-01 6.930e-02 -5.192 2.37e-07 ***
## Site_recCISSIN_2020_SITE_G -2.272e+00 1.251e-01 -18.166 < 2e-16 ***
## Site_recOUAGA_2000_SITE_A   1.883e-01 4.461e-02  4.222 2.56e-05 ***
## Site_recSECTEUR_16_OUAGA    -4.912e-01 2.395e-01 -2.051 0.040463 *
## Site_recSILMIOUGOU          2.624e-01 5.344e-02  4.911 1.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2322 on 1480 degrees of freedom
## Multiple R-squared:  0.8956, Adjusted R-squared:  0.8938
## F-statistic: 488.5 on 26 and 1480 DF,  p-value: < 2.2e-16

```

Ainsi donc le meilleur modèle reste le modèle que nous avons ajuster

3. Test d'hypthèses

a. Autocorrelation

```

dwtest(modele_selectionne)

##
## Durbin-Watson test
##
## data: modele_selectionne
## DW = 0.96083, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

check_autocorrelation(modele_selectionne)

## Warning: Autocorrelated residuals detected (p < .001).

```

Le test Durbin-Watson indique que les résidus sont significativement autocorrélés positivement (avec une valeur DW de 1.0357 et une p-value très basse). Cela signifie que les erreurs du modèle ne sont pas indépendantes les unes des autres.

b. Homoscedasticité

```
bptest(modlm)#Test de Breusch-Pagan

##
## studentized Breusch-Pagan test
##
## data: modlm
## BP = 661.02, df = 26, p-value < 2.2e-16

check_heteroscedasticity(modele_selectionne)

## Warning: Heteroscedasticity (non-constant error variance) detected (p < .001).
```

le test de Breusch-Pagan détecte une hétérosécédasticité des residus ce qui indique que la variance des erreurs est non constante.

c. Normalité des residus

```
shapiro.test(residuals(modele_selectionne))

##
## Shapiro-Wilk normality test
##
## data: residuals(modele_selectionne)
## W = 0.87801, p-value < 2.2e-16

check_normality(modele_selectionne)

## Warning: Non-normality of residuals detected (p < .001).
```

Le test de Shapiro-Wilk révèle que les résidus du modèle ne sont pas normalement distribués.

d. Colinearité

```
check_collinearity(modele_selectionne)

## # Check for Multicollinearity
##
## Low Correlation
##
##             Term  VIF      VIF 95% CI Increased SE Tolerance Tolerance
## 95% CI
##          Superficie 2.02 [ 1.88,   2.18]       1.42      0.50 [0.46,
## 0.53]
##          Cout_m2 4.81 [ 4.40,   5.26]       2.19      0.21 [0.19,
## 0.23]
##  Taxe_Jouissance 3.14 [ 2.89,   3.42]       1.77      0.32 [0.29,
```

```

0.35]
##
## Moderate Correlation
##
##           Term   VIF      VIF 95% CI Increased SE Tolerance Tolerance 95%
CI
##  plan_etablie 8.66 [ 7.89,   9.51]          2.94     0.12     [0.11, 0.
13]
##
## High Correlation
##
##           Term   VIF      VIF 95% CI Increased SE Tolerance
##       Usage 17.98 [ 16.34,  19.81]        4.24     0.06
##  attestation_etablie 11.32 [ 10.31,  12.45]        3.37     0.09
##       years 38.25 [ 34.69,  42.18]        6.18     0.03
##      Site_rec 133.97 [121.39, 147.86]       11.57 7.46e-03
## Tolerance 95% CI
##      [0.05, 0.06]
##      [0.08, 0.10]
##      [0.02, 0.03]
##      [0.01, 0.01]

```

Les variable Superficie, Taxe_Jouissance, et attestation_etablie présente des multicolinéarité faible et modérée ce qui nous indique une stabilité de ces variable. cependant les autre variables comme Usage, years et Site_rec ont une multicolinéarité élevée ce qui rend les estimation peu fiables.

e. Conclusion sur les hypothèses

Nous constatons que les principaux tests d'hypothèses ne sont pas vérifiés sur le modèle de régression linéaire multiple que nous venons d'ajuster. En conséquence, notre modèle ne s'adapte pas adéquatement à nos données. Pour remédier à ce problème, nous allons utiliser le modèle de régression LASSO, qui nous permettra de pallier ces insuffisances

VI. Régression Lasso

Nous pouvons utiliser le modèle LASSO (Least Absolute Shrinkage and Selection Operator) pour estimer les prix hédoniques. Le modèle LASSO est une méthode de régularisation qui peut aider à améliorer les prédictions en réduisant le surajustement et en sélectionnant les variables les plus importantes. La sélection des variables à inclure ou à exclure du modèle nécessite une analyse approfondie de la théorie économique derrière la demande de parcelles. Cependant, si la question de recherche implique d'avoir un modèle simple, avec la plus petite erreur et une meilleure précision prédictive, nous utilisons la régression Lasso de l'apprentissage automatique pour éviter cette tâche. Considérez le Lasso comme une régression OLS qui impose une pénalité (λ) pour la taille des estimations des paramètres. Le paramètre λ est la pénalité de régularisation. Lorsque λ tend vers zéro, nous obtenons les mêmes estimations que dans la régression OLS, et lorsqu'il tend vers l'infini, nous obtenons un modèle avec des estimations nulles. Par défaut, la fonction *glmnet()* effectue une régression Lasso pour une gamme automatiquement sélectionnée de valeurs de λ . Cependant, ici nous avons choisi d'implémenter la fonction sur une grille de valeurs allant de $\lambda = 10^{10}$ à $\lambda =$

10^{-2} , couvrant essentiellement toute la gamme de scénarios, du modèle nul contenant uniquement l'ordonnée à l'origine, à l'ajustement par moindres carrés. Ce code nous permet d'utiliser le modèle LASSO pour estimer les prix hédoniques, en prenant en compte la régularisation pour améliorer les prédictions et sélectionner les variables les plus pertinentes.

```
grid = 10^seq(10, -2, length = 100)
```

1. Configuration des données

```
# supprimer La première colonne et ne Laisser que Les variables indépendantes
x = model.matrix(lnCOUT~Usage + Superficie + Cout_m2 + Taxe_Jouissance + attestation_etablie + plan_etablie + years + Site_rec, hdnc)[,-1]
# ne sélectionner que La variable dépendante
y = hdnc %>% select(lnCOUT) %>% unlist() %>% as.numeric()
```

Nous allons maintenant diviser les échantillons en un ensemble d'entraînement et un ensemble de test afin d'estimer l'erreur de test de la régression lasso.

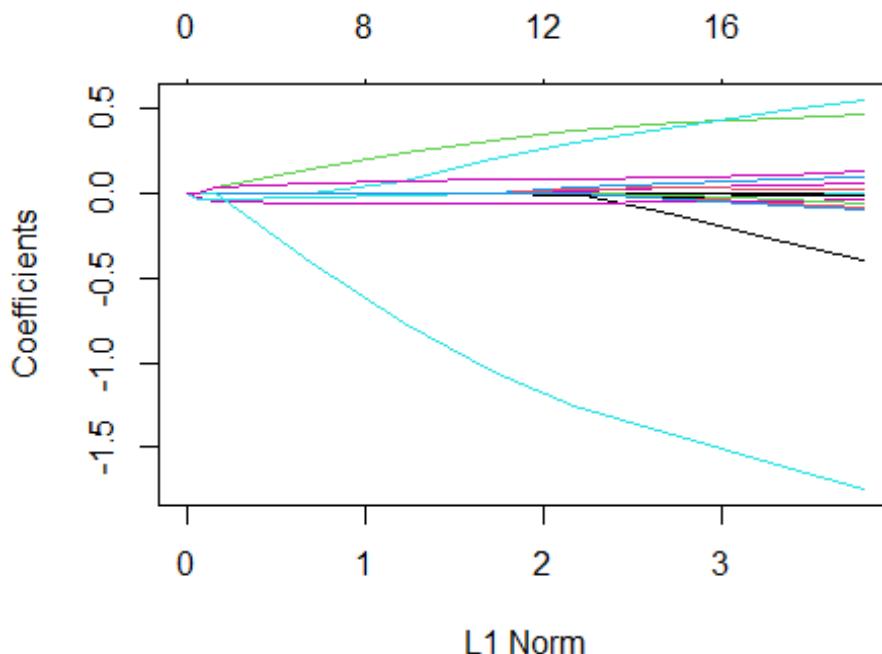
```
set.seed(1)
# nous sélectionnons La moitié des données pour entraîner Le modèle et L'autre moitié pour le tester
train = hdnc %>% sample_frac(0.5)
test = hdnc %>% setdiff(train)
# nous définissons maintenant Les variables dépendantes et indépendantes
x_train = model.matrix(lnCOUT~Usage + Superficie + Cout_m2 + Taxe_Jouissance + attestation_etablie + plan_etablie + years + Site_rec, train)[,-1]
x_test = model.matrix(lnCOUT~Usage + Superficie + Cout_m2 + Taxe_Jouissance + attestation_etablie + plan_etablie + years + Site_rec, test)[,-1]
y_train = train %>% select(lnCOUT) %>% unlist() %>% as.numeric()
y_test = test %>% select(lnCOUT) %>% unlist() %>% as.numeric()
```

Associé à chaque valeur de λ se trouve un vecteur de coefficients de régression ridge, stocké dans une matrice accessible via la fonction `coef()`. Dans ce cas, il s'agit d'une matrice de 27×100 , avec 27 lignes (une pour chaque prédicteur, plus une constante) et 100 colonnes (une pour chaque valeur de λ).

La fonction `glmnet()` a un argument `alpha` qui détermine le type de modèle ajusté. Si `alpha = 0`, un modèle de régression ridge est ajusté, et si `alpha = 1`, un modèle lasso est ajusté

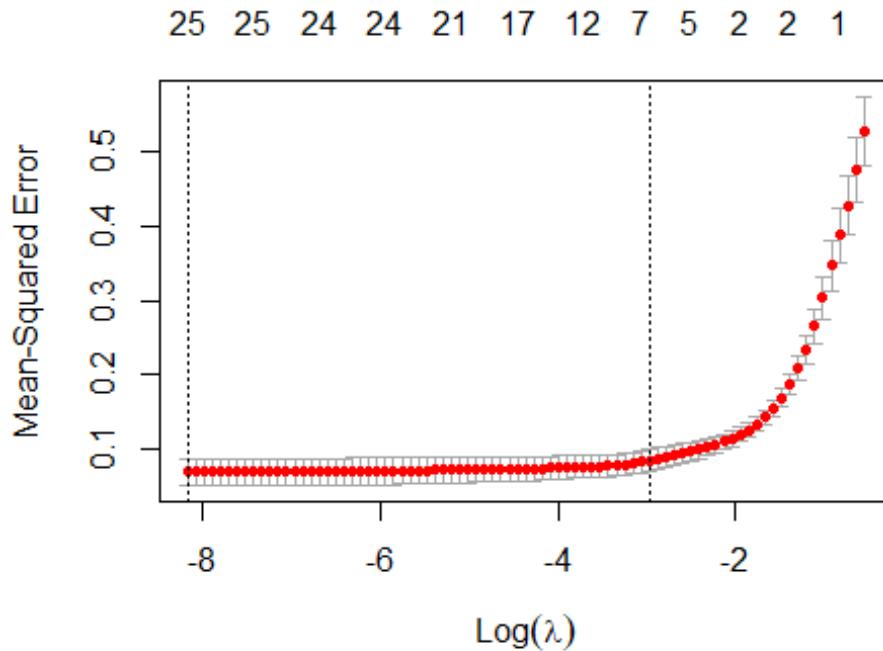
```
lasso_mod = glmnet(x_train, y_train, alpha = 1, lambda = grid)# Ajuster Le modèle Lasso sur Les données d'entraînement
dim(coef(lasso_mod))# Dimensions de la matrice des coefficients
## [1] 27 100
plot(lasso_mod)# Tracer Le graphique des coefficients
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## suppression des ex-aequos de 'x'
```

2. Les graphiques des coefficients



On remarque que dans le graphique des coefficients, en fonction du choix du paramètre de régularisation, certains coefficients sont exactement égaux à zéro. Nous effectuons maintenant une validation croisée (test hors échantillon) et calculons l'erreur de test associée :

```
set.seed(1)
cv.out = cv.glmnet(x_train, y_train, alpha = 1) # Ajuster le modèle Lasso sur les données d'entraînement
plot(cv.out) # Tracer le graphique de l'erreur quadratique moyenne (MSE) de l'entraînement en fonction de Lambda
```



```
bestlam = cv.out$lambda.min # Sélectionner le Lambda qui minimise l'erreur
# quadratique moyenne (MSE) sur les données d'entraînement
lasso_pred = predict(lasso_mod, s = bestlam, newx = x_test) # Utiliser le meilleur Lambda pour prédire les données de test
mean((lasso_pred - y_test)^2) # Calculate test MSE # Calculer l'erreur quadratique moyenne (MSE) sur les données de test
## [1] 0.1337675
```

Ceci est considérablement inférieur à la MSE de l'ensemble de test du modèle nul et de la méthode des moindres carrés. De plus, le lasso présente un avantage significatif par rapport aux autres régressions en ce que les estimations des coefficients résultants sont clairsemées. Ici, nous voyons qu'un grand nombre des 27 estimations de coefficients sont exactement égales à zéro :

3. Ajuster le modèle LASSO sur l'ensemble des données

```
out = glmnet(x, y, alpha = 1, lambda = grid) # Ajuster le modèle Lasso sur
# l'ensemble des données
lasso_coef = predict(out, type = "coefficients", s = bestlam)[1:27,] # Afficher les coefficients en utilisant le Lambda choisi par la validation croisée
lasso_coef
## (Intercept) UsageCOMMERCE A L'ANGLE
## 1.502292e+01 0.000000e+00
## UsageCOMMERCE ANGLE UsageCOMMERCE ANGLE 1 BITUME
## -3.360304e-01 -4.491364e-02
## UsageCOMMERCE ANGLE 2 VOIES UsageCOMMERCE ORDINAIRE ANGLE
## 3.375658e-01 1.353208e-01
## UsageCOMMUNAUTAIRE UsageHABITATION
```

```

##          1.427123e-01          -1.507675e-01
## UsageHABITATION ANGLE      UsageSTATION SERVICE
##          -1.340192e-01          2.617122e-02
##           Superficie           Cout_m2
##          3.802249e-04          3.612312e-05
## Taxe_Jouissance attestatIOn etablieNON DEFINI
##          6.955596e-05          0.000000e+00
## attestatIOn etablieOUI      plan_etablieNON DEFINI
##          3.903457e-02          0.000000e+00
## plan_etablieOUI            years2019
##          0.000000e+00          1.766106e-02
## years2020                  years2021
##          -4.432560e-02         -1.131131e-01
## years2022                  years2023
##          0.000000e+00          0.000000e+00
## years2024                  Site_recCISSIN_2020_SITE_G
##          0.000000e+00          -1.667337e+00
## Site_recOUAGA_2000_SITE_A  Site_recSECTEUR_16_OUAGA
##          1.316929e-01         -2.229008e-01
## Site_recSILMIOUGOU        0.000000e+00
##
```

En sélectionnant uniquement les prédicteurs avec des coefficients non nuls, nous voyons que le modèle lasso avec λ choisi par validation croisée ne contient que 18 variables :

```



lasso_coef[lasso_coef != 0] # Afficher uniquement les coefficients non nuls



```

(Intercept) UsageCOMMERC E ANGLE
1.502292e+01 -3.360304e-01
UsageCOMMERC E ANGLE 1 BITUME UsageCOMMERC E ANGLE 2 VOIES
-4.491364e-02 3.375658e-01
UsageCOMMERC E ORDINAIRE ANGLE UsageCOMMUNAUTAIRE
1.353208e-01 1.427123e-01
UsageHABITAT ION UsageHABITAT ION ANGLE
-1.507675e-01 -1.340192e-01
UsageSTATION SERVICE Superficie
2.617122e-02 3.802249e-04
##
```


```

```

##                  Cout_m2          Taxe_Jouissance
##                  3.612312e-05    6.955596e-05
##      attestation_etablieOUI      years2019
##                  3.903457e-02    1.766106e-02
##      years2020                  years2021
##                  -4.432560e-02   -1.131131e-01
##      Site_recCISSIN_2020_SITE_G Site_recOUAGA_2000_SITE_A
##                  -1.667337e+00   1.316929e-01
##      Site_recSECTEUR_16_OUAGA
##                  -2.229008e-01

```

4. Interprétation des Estimations du Logarithme du Coût des Parcelles

Les coefficients de ce modèle fournissent des informations sur l'effet marginal de chaque variable explicative sur le logarithme du coût des parcelles, en tenant compte des autres variables. Les signes et les magnitudes des coefficients permettent de comprendre comment chaque caractéristique affecte la variable dépendante. Par exemple, certains usages, années et sites ont des effets positifs ou négatifs significatifs sur la variable dépendante. En effet tout choses étant égales par ailleurs : Voici une version révisée et améliorée du texte concernant l'estimation du logarithme du coût des parcelles par le modèle LASSO :

1. Parcelles à Usage Commercial à l'Angle :

- Une augmentation d'une unité des parcelles à usage commercial situées à l'angle réduit le logarithme du coût des parcelles de 0,336.

2. Parcelles à Usage Commercial à l'Angle d'une Route Bitumée :

- Une augmentation d'une unité des parcelles à usage commercial situées à l'angle d'une route bitumée diminue le logarithme du coût des parcelles de 0,449.

3. Parcelles à Usage Commercial à l'Angle d'une Route à Deux Voies :

- Une augmentation d'une unité des parcelles à usage commercial situées à l'angle d'une route à deux voies augmente le logarithme du coût des parcelles de 0,336.

4. Parcelles à Usage Commercial Ordinaire à l'Angle :

- Une augmentation d'une unité des parcelles à usage commercial ordinaire situées à l'angle accroît le logarithme du coût des parcelles de 0,135.

5. Parcelles à Usage Communautaire :

- Une augmentation d'une unité des parcelles à usage communautaire augmente le logarithme du coût des parcelles de 0,143.

6. Parcelles à Usage d'Habitation :

- Une augmentation d'une unité des parcelles à usage d'habitation diminue le logarithme du coût des parcelles de 0,1508.

7. Parcelles à Usage d'Habitation à l'Angle :

- Une augmentation d'une unité des parcelles à usage d'habitation situées à l'angle réduit le logarithme du coût des parcelles de 0,1340.

8. Stations-Service :

- Une augmentation d'une unité du nombre de stations-service augmente le logarithme du coût des parcelles de 0,0262.

9. Superficie :

- Pour chaque unité supplémentaire de superficie, le logarithme du coût des parcelles augmente de 0,0003802.

10. Coût par Mètre Carré :

- Pour chaque unité supplémentaire du coût par mètre carré, le logarithme du coût des parcelles augmente de 0,00003612.

11. Taxe de Jouissance :

- Pour chaque unité supplémentaire de taxe de jouissance, le logarithme du coût des parcelles augmente de 0,00006956.

12. Attestation Établie :

- Le fait que l'attestation soit établie augmente le logarithme du coût des parcelles de 0,0390.

13. Plan Établi :

- Le fait que le plan ne soit pas défini n'a pas d'effet direct sur le logarithme du coût des parcelles ; cet effet est pris comme référence.

14. Années :

- En 2019, le logarithme du coût des parcelles augmente de 0,0177 par rapport à l'année de référence.
- En 2020, le logarithme du coût des parcelles diminue de 0,0443 par rapport à l'année de référence.
- En 2021, le logarithme du coût des parcelles diminue de 0,1131 par rapport à l'année de référence.

15. Sites :

- Le site CISSIN en 2020 (SITE_G) réduit le logarithme du coût des parcelles de 1,6673.
- Le site OUAGA_2000 (SITE_A) augmente le logarithme du coût des parcelles de 0,1317.
- Le secteur 16 de OUAGA réduit le logarithme du coût des parcelles de 0,2229.
- Le site SILMIOUGOU n'a pas d'effet direct sur le logarithme du coût des parcelles

estimateurs sans direct effet direct sur la le logarithme du coût des parcelles

Les années 2022, 2023 et 2024 n'ont pas d'effet direct sur la variable dépendante les parcelles qui ont un plan_etablie tout comme les parcelles à plan NON DEFINI, les parcelles à Usage COMMERCiale A L'ANGLE, les parcelles qui ont des attestation non definitivement etablie, ont pas d'effets direct sur le le logarithme du coût des parcelles

Obtenir les coefficients en utilisant le lambda optimal

```
best_lasso_coef = predict(out, type = "coefficients", s = bestlam)[,1]
```

Préparer les matrices de caractéristiques pour toutes les années

```
x_all = model.matrix(lnCOUT~Usage + Superficie + Cout_m2 + Taxe_Jouissance
+ attestation_etablie + plan_etablie + years + Site_rec, hdnc)
```

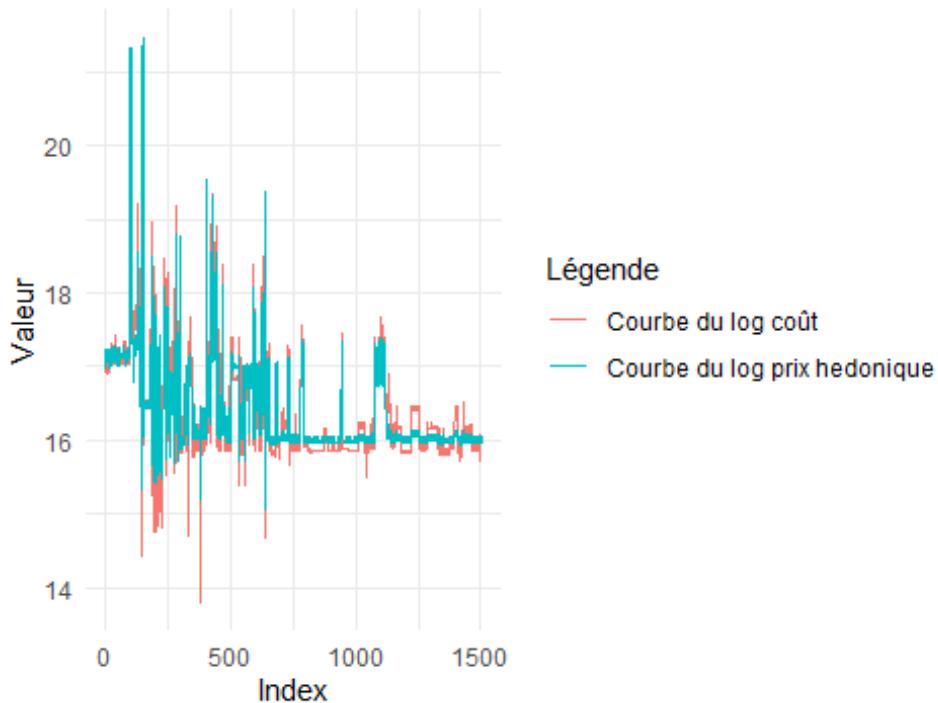
5. Prédire les valeurs de lnCOUT pour toutes les années

```
hdnc$hedonic_lnCOUT = x_all %*% best_lasso_coef
```

Représentation de courbe du log coût et de la courbe du log prix hedonique

```
library(ggplot2)
hdnc$index <- 1:nrow(hdnc)
ggplot(hdnc, aes(x = index)) +
  geom_line(aes(y = lnCOUT, color = "Courbe du log coût")) +
  geom_line(aes(y = hedonic_lnCOUT, color = "Courbe du log prix hedonique"))
) +
  labs(title = "Représentation des courbes",
       x = "Index",
       y = "Valeur",
       color = "Légende") +
  theme_minimal()
```

Représentation des courbes



6. Les indices élémentaires ou simples

Convertir en coût et Calculer les Indices :

```
hdnc$hedonic_COST = exp(hdnc$hedonic_lnCOUT)

# Calculer le prix moyen, la superficie moyenne, pour l'année
indices = hdnc %>%
  group_by(years) %>%
  summarise(COUT_total = mean(lnCOUT), area = mean(Superficie), PrixH = mea
n(hedonic_COST), )

PrixH_2018 <- indices$PrixH[indices$years == "2018"]
PrixH_2019 <- indices$PrixH[indices$years == "2019"]
PrixH_2020 <- indices$PrixH[indices$years == "2020"]
PrixH_2021 <- indices$PrixH[indices$years == "2021"]
PrixH_2022 <- indices$PrixH[indices$years == "2022"]
```

```

PrixH_2023 <- indices$PrixH[indices$years == "2023"]
PrixH_2024 <- indices$PrixH[indices$years == "2024"]

Ih18.19 <- (PrixH_2019 / PrixH_2018) * 100
Ih18.20 <- (PrixH_2020 / PrixH_2018) * 100
Ih18.21 <- (PrixH_2021 / PrixH_2018) * 100
Ih18.22 <- (PrixH_2022 / PrixH_2018) * 100
Ih18.23 <- (PrixH_2023 / PrixH_2018) * 100
Ih18.24 <- (PrixH_2024 / PrixH_2018) * 100

# Afficher L'indice de prix hédonique

#L'indice de prix hédonique de 2019 par rapport à 2018 est de
round(Ih18.19, 2)
## [1] 198.34

#L'indice de prix hédonique de 2020 par rapport à 2018 est de
round(Ih18.20, 2)
## [1] 108.14

#L'indice de prix hédonique de 2021 par rapport à 2018 est de
round(Ih18.21, 2)
## [1] 609

#L'indice de prix hédonique de 2022 par rapport à 2018 est de
round(Ih18.22, 2)
## [1] 150.34

#L'indice de prix hédonique de 2023 par rapport à 2018 est de
round(Ih18.23, 2)
## [1] 50.37

#L'indice de prix hédonique de 2024 par rapport à 2018 est de
round(Ih18.24, 2)
## [1] 48.01

```

a. Interpretation des indices des prix

Les données révèlent une forte volatilité des prix des parcelles de terrain à Ouagadougou entre 2018 et 2024. En particulier, la hausse spectaculaire des prix en 2019 et surtout en 2021 indique des périodes de forte demande ou des événements spécifiques ayant influencé le marché immobilier. Cette flambée pourrait être attribuée à des facteurs tels qu'une demande accrue, des investissements majeurs ou des changements dans les conditions économiques locales. En revanche, les baisses successives observées en 2023 et 2024 suggèrent un possible retournement du marché, possiblement en réponse à des évolutions économiques, sociales ou politiques défavorables. Ces indices sont cruciaux pour comprendre les dynamiques du marché immobilier et peuvent guider la formulation de politiques urbaines et économiques adaptées aux fluctuations des prix du foncier.

Taux de croissances

```
t19 = Ih18.19/100-1
t20 = Ih18.20/100-1
t21 = Ih18.21/100-1
t22 = Ih18.22/100-1
t23 = Ih18.23/100-1
t24 = Ih18.24/100-1

#Afficher taux de croissance

#taux de croissance de 2019 par rapport à 2018 est de
round(t19, 2)
## [1] 0.98

#taux de croissance de 2020 par rapport à 2018 est de
round(t20, 2)
## [1] 0.08

#taux de croissance de 2021 par rapport à 2018 est de
round(t21, 2)
## [1] 5.09

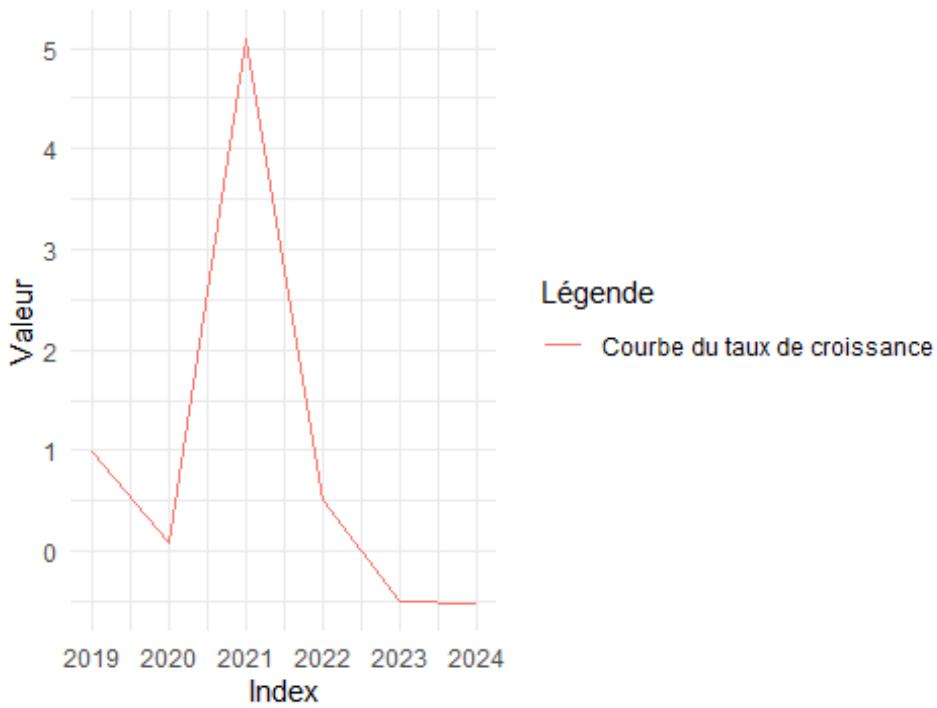
#taux de croissance de 2022 par rapport à 2018 est de
round(t22, 2)
## [1] 0.5

#taux de croissance de 2023 par rapport à 2018 est de
round(t23, 2)
## [1] -0.5

#taux de croissance de 2024 par rapport à 2018 est de
round(t24, 2)
## [1] -0.52

t = c(t19,t20,t21,t22,t23,t24)
t = as.data.frame(t)
t$index <- c(2019,2020,2021,2022,2023,2024)
ggplot(t, aes(x = index)) +
  geom_line(aes(y = t, color = "Courbe du taux de croissance")) +
  labs(title = "Représentation de deux courbes",
       x = "Index",
       y = "Valeur",
       color = "Légende") +
  theme_minimal()
```

Représentation de deux courbes



moyenne par année

```
area_2018 <- indices$area[indices$years == "2018"]
area_2019 <- indices$area[indices$years == "2019"]
area_2020 <- indices$area[indices$years == "2020"]
area_2021 <- indices$area[indices$years == "2021"]
area_2022 <- indices$area[indices$years == "2022"]
area_2023 <- indices$area[indices$years == "2023"]
area_2024 <- indices$area[indices$years == "2024"]
```

#L'indice de Valeur

```
v19=(PrixH_2019*area_2019 / (PrixH_2018*area_2018))*100
v20=(PrixH_2020*area_2020 / (PrixH_2019*area_2019))*100
v21=(PrixH_2021*area_2021 / (PrixH_2020*area_2020))*100
v22=(PrixH_2022*area_2022 / (PrixH_2021*area_2021))*100
v23=(PrixH_2023*area_2023 / (PrixH_2022*area_2022))*100
v24=(PrixH_2024*area_2024 / (PrixH_2023*area_2023))*100
v= c(v19, v20, v21, v22, v23, v24)
```

#Afficher l'indice de Valeur

```
#L'indice de Valeur de 2019 par rapport à 2018 est de
round(Ih18.19, 2)
```

```
## [1] 198.34
```

```
#L'indice de Valeur de 2020 par rapport à 2018 est de
round(Ih18.20, 2)
```

```
## [1] 108.14
```

la superficie

```

#L'indice de Valeur de 2021 par rapport à 2018 est de
round(Ih18.21, 2)

## [1] 609

#L'indice de Valeur de 2022 par rapport à 2018 est de
round(Ih18.22, 2)

## [1] 150.34

#L'indice de Valeur de 2023 par rapport à 2018 est de
round(Ih18.23, 2)

## [1] 50.37

#L'indice de Valeur de 2024 par rapport à 2018 est de
round(Ih18.24, 2)

## [1] 48.01

```

b. Interpretraion des indices de valeurs

Les indices de valeur révèlent des variations importantes des prix des parcelles de terrain, similaires à celles observées avec les indices hédoniques et de Laspeyres. La forte hausse des prix en 2021 est particulièrement remarquable, suggérant des conditions économiques exceptionnelles ou des changements significatifs dans le marché immobilier. La tendance à la baisse des prix en 2023 et 2024 indique un possible ajustement du marché ou des réponses à des facteurs externes, tels que des changements économiques ou des politiques défavorables. En utilisant les prix de la période de base et les quantités actuelles, les indices de valeur offrent une perspective complémentaire sur les dynamiques du marché immobilier, permettant une analyse plus complète et la formulation de politiques appropriées.

#L'indice de Laspeyres

```

119=(PrixH_2019*area_2018 / (PrixH_2018*area_2018))*100
120=(PrixH_2020*area_2018 / (PrixH_2018*area_2018))*100
121=(PrixH_2021*area_2018 / (PrixH_2018*area_2018))*100
122=(PrixH_2022*area_2018 / (PrixH_2018*area_2018))*100
123=(PrixH_2023*area_2018 / (PrixH_2018*area_2018))*100
124=(PrixH_2024*area_2018 / (PrixH_2018*area_2018))*100
l= c(119, 120, 121, 122, 123, 124)

```

Afficher l'indice de Laspeyres

```

#L'indice de Laspeyres de 2019 par rapport à 2018 est de
round(119, 2)

## [1] 198.34

#L'indice de Laspeyres de 2020 par rapport à 2018 est de
round(120, 2)

## [1] 108.14

```

```

#L'indice de Laspeyres de 2021 par rapport à 2018 est de
round(121, 2)

## [1] 609

#L'indice de Laspeyres de 2022 par rapport à 2018 est de
round(122, 2)

## [1] 150.34

#L'indice de Laspeyres de 2023 par rapport à 2018 est de
round(123, 2)

## [1] 50.37

#L'indice de Laspeyres de 2024 par rapport à 2018 est de
round(124, 2)

## [1] 48.01

```

c. Interprétation des indices de Laspeyres

Les indices de Laspeyres révèlent une volatilité des prix similaire à celle observée avec les indices hédoniques, soulignant des fluctuations importantes dans le marché immobilier. La forte augmentation des prix en 2021 est particulièrement marquante, suggérant des conditions économiques exceptionnelles ou des changements significatifs dans le marché. La tendance à la baisse des prix en 2023 et 2024 pourrait indiquer un ajustement du marché ou des réponses à des facteurs externes, tels que des politiques économiques ou des conditions défavorables. Comparés aux indices hédoniques, les indices de Laspeyres offrent une perspective distincte sur les changements de prix en utilisant les quantités de la période de base, fournissant ainsi des informations complémentaires pour comprendre les dynamiques du marché immobilier et orienter les décisions politiques et économiques.

#L'indice de Paasche

```

p19=(PrixH_2019*area_2019 / (PrixH_2019*area_2018))*100
p20=(PrixH_2020*area_2020 / (PrixH_2020*area_2018))*100
p21=(PrixH_2021*area_2021 / (PrixH_2021*area_2018))*100
p22=(PrixH_2022*area_2022 / (PrixH_2022*area_2018))*100
p23=(PrixH_2023*area_2023 / (PrixH_2023*area_2018))*100
p24=(PrixH_2024*area_2024 / (PrixH_2024*area_2018))*100
p= c(p19, p20, p21, p22, p23, p24)

```

Afficher l'indice de Laspeyres

```

#L'indice de Paasche de 2019 par rapport à 2018 est de
round(p19, 2)

## [1] 116.21

#L'indice de Paasche de 2020 par rapport à 2018 est de
round(p20, 2)

## [1] 100.28

```

```

#L'indice de Paasche de 2021 par rapport à 2018 est de
round(p21, 2)

## [1] 183.03

#L'indice de Paasche de 2022 par rapport à 2018 est de
round(p22, 2)

## [1] 139.3

#L'indice de Paasche de 2023 par rapport à 2018 est de
round(p23, 2)

## [1] 54.47

#L'indice de Paasche de 2024 par rapport à 2018 est de
round(p24, 2)

## [1] 52.73

```

d. Interpretation des indices de Paasche

Les indices de Paasche révèlent des variations des prix qui intègrent les quantités de chaque année, offrant ainsi une perspective distincte par rapport aux indices de Laspeyres et de valeur. La forte augmentation des prix en 2021 est particulièrement marquante, suggérant des conditions exceptionnelles ou des changements significatifs sur le marché immobilier. La tendance à la baisse des prix en 2023 et 2024 est également notable, indiquant un ajustement du marché ou des réponses à des facteurs externes. En utilisant les quantités de la période en cours, les indices de Paasche enrichissent l'analyse globale des dynamiques des prix et facilitent la formulation de politiques adaptées aux variations du marché immobilier.

#l'indice de Fisher

```

f19 = sqrt(p19*l19)
f20 = sqrt(p20*l20)
f21 = sqrt(p21*l22)
f22 = sqrt(p22*l22)
f23 = sqrt(p23*l23)
f24 = sqrt(p24*l24)
f = c(f19,f20,f21,f22,f23,f24)

# Afficher l'indice de Fisher
#L'indice de Fisher de 2019 par rapport à 2018 est de
round(f19, 2)

## [1] 151.82

#L'indice de Fisher de 2020 par rapport à 2018 est de
round(f20, 2)

## [1] 104.14

#L'indice de Fisher de 2021 par rapport à 2018 est de
round(f21, 2)

```

```

## [1] 165.88
#L'indice de Fisher de 2022 par rapport à 2018 est de
round(f22, 2)

## [1] 144.72
#L'indice de Fisher de 2023 par rapport à 2018 est de
round(f23, 2)

## [1] 52.38
#L'indice de Fisher de 2024 par rapport à 2018 est de
round(f24, 2)

## [1] 50.31

```

e. Interprétation des indices de Fischer

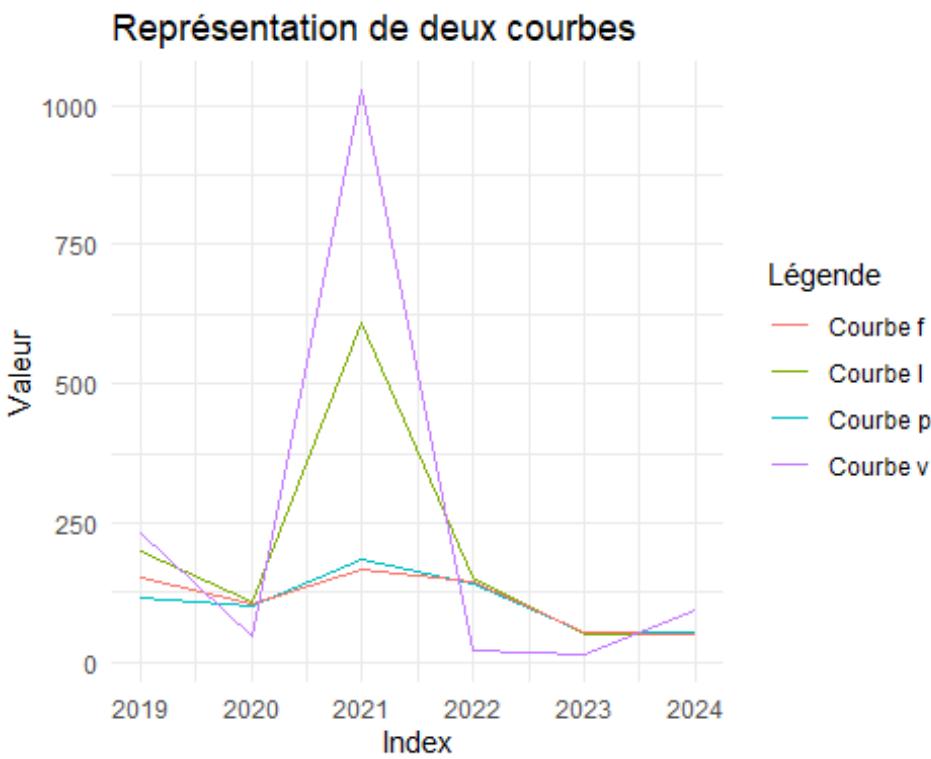
Les indices de Fisher montrent des variations des prix qui intègrent les perspectives des indices de Laspeyres et de Paasche, offrant ainsi une vue plus équilibrée sur les changements de prix. La forte augmentation des prix en 2021 est particulièrement notable, suggérant des conditions exceptionnelles ou des changements significatifs dans le marché immobilier. La tendance à la baisse des prix en 2023 et 2024 est également marquante, indiquant un ajustement ou une réponse à des facteurs externes. En fournissant une mesure intégrée des variations des prix, les indices de Fisher enrichissent l'analyse globale du marché immobilier et facilitent la formulation de politiques adaptées aux fluctuations du marché.

#représentation graphique

```

plott= data.frame(l,p,f,v)
plott$index = c(2019,2020,2021,2022,2023,2024)
ggplot(plott, aes(x = index)) +
  geom_line(aes(y = l, color = "Courbe l")) +
  geom_line(aes(y = p, color = "Courbe p")) +
  geom_line(aes(y = f, color = "Courbe f")) +
  geom_line(aes(y = v, color = "Courbe v")) +
  labs(title = "Représentation de deux courbes",
       x = "Index",
       y = "Valeur",
       color = "Légende") +
  theme_minimal()

```



Conclusion

L'analyse avec le modèle hédonique révèle que la superficie, la localisation, l'usage et l'année de vente sont des facteurs clés influençant les prix des terrains. Cette décomposition fournit des informations essentielles pour orienter les politiques publiques et les investissements urbains futurs.