Université Joseph Ki-Zerbo
Institut Supérieur des Sciences de la Population
Licence Professionnelle en Analyse Statistique - 2e année

Économétrie des variables Quantitatives

Modélisation de l'indice du prix des parcelle entre 2018 et 2024 dans la ville de Ouagadougou

Réalisé par :
NIAMPA Abdoul Fataho
SAWADOGO Pengdwendé Orianne-Aurele
YAMEOGO Saïdou

Enseignant :
Dr. Boyam Fabrice YAMEOGO

Juin 2025

Table des matières

0.1 Introduction

Le développement urbain rapide que connaissent de nombreuses villes africaines s'accompagne d'une pression croissante sur les ressources foncières. À Ouagadougou, la capitale du Burkina Faso, cette dynamique se traduit par une intensification des transactions foncières et une augmentation progressive du coût des parcelles. Face à cette évolution, il devient crucial pour les acteurs publics et privés, notamment les aménageurs et les autorités locales, de disposer d'outils de mesure fiables permettant de suivre l'évolution des prix du foncier.

L'évaluation de l'évolution des prix des parcelles immobilières constitue un enjeu majeur pour les acteurs économiques, les décideurs politiques et les investisseurs, permettant de mieux comprendre les dynamiques du marché immobilier et d'orienter les stratégies d'investissement ou de régulation. Dans ce contexte, la période s'étendant de 2018 à 2024 offre un cadre pertinent pour analyser les fluctuations des prix, marquées par des événements économiques et sociaux significatifs.

Cependant, la construction d'un tel indicateur pose plusieurs défis. Une simple comparaison des prix moyens d'une année à l'autre ne permet pas de distinguer les variations réelles des prix de celles liées à des changements dans la nature des parcelles vendues (localisation, superficie, usage, statut administratif, etc.). Autrement dit, il est essentiel de neutraliser l'effet qualité pour isoler la variation purement temporelle des prix.

C'est dans ce cadre que s'inscrit la méthode hédonique, utilisée dans de nombreux pays pour la construction d'indices immobiliers ajustés. Cette méthode repose sur l'idée que le prix d'un bien est fonction de ses caractéristiques observables. Elle permet ainsi de modéliser la contribution individuelle de chaque attribut (par exemple : superficie, quartier, usage prévu) à la formation du prix, et de construire un indice de prix corrigé des effets de composition.

0.2 Présentation des données

0.2.1 Sources de données

Les données utilisées dans cette étude proviennent de la Société Nationale d'Aménagement des Terrains Urbains (SONATUR). Elles portent sur l'ensemble des parcelles vendues à Ouagadougou

entre 2018 et 2024, issues de différents sites d'aménagement. Chaque enregistrement correspond à une parcelle individuelle pour laquelle diverses caractéristiques ont été renseignées.

La base de données comprend 1811 d'observations et 15 variables et couvre aussi bien des zones périphériques que des quartiers plus centraux. Elle constitue une source précieuse pour analyser l'évolution des prix fonciers et modéliser la valeur des terrains à partir de leurs attributs.

0.2.2 Analyse descriptive préliminaire

Avant de procéder à la construction des variables synthétiques et à la modélisation, une analyse descriptive des données a été réalisée pour explorer les caractéristiques des parcelles et identifier les relations potentielles entre les variables.

0.2.2.1 - Variables quantitatives Le tableau ci-dessous présente les indicateurs statistiques clés moyenne, médiane, écart-type, minimum, maximum et asymétrie pour variables quantitatives : le cout par m2 (Cout_m2),le coût total de parcelles (COUT), la superficie (superficie) et le taxe de jouissance (Taxe_Jouissance).

Table 1: Résumé statistique des variables quantitatives

Variable	Moyenne	Mediane	Ecart_type	Minimum	Maximum	Asymetrie
Cout_m2	27349.66	26000	17584.16	0	190000	1.68
COUT	16068350.47	8320000	35286959.09	0	722376000	11.13
Superficie	495.01	320	720.28	82	13611	9.84
Taxe_Jouissance	514.77	500	492.00	0	3000	2.89

Le résumé statistique des variables quantitatives met en évidence des disparités notables entre les indicateurs observés. Le coût par m2 (Cout_m2), le coût total (COUT) et les taxes de jouissance (Taxe_Jouissance) présente une moyenne élevée respectivement 27349.66 FCFA, 16068350.47 FCFA, 514.77 FCFA, mais une forte asymétrie positive respectivement 1.68, 11.13, 2.89 indiquant la présence de valeurs extrêmes vers le haut, ce que confirme l'écart important entre le minimum respectivement 0, 0, 0 et le maximum respectivement 190000 FCFA, 722376000 FCFA,3000 FCFA. Quant à la superficie, on remarque également une forte asémetrie (9,84) temoignant un étalement vers la droite.

On note une incohérence dans les données via ces variables: des parcelles qui coûtent 0 FCFA.

0.2.2.2 - Variables qualitatives

Table 2: Résumé descriptif des variables qualitatives

Variable	Nb_modalites	Modalite_dominante	Effectif	Pourcentage	Valeurs_manquantes
Ville	1	OUAGADOUGOU	1811	100	0
Site	6	SILMIOUGOU	1170	64.6	0
Usage	10	HABITATION	1169	64.5	0
Type_option	3	ACOMPTE 30%	1365	75.4	1
attestation_etablie	3	NON DEFINI	1306	72.1	0
plan_etablie	3	NON DEFINI	1358	75	0
Presence_ONEA	1	OUI	1811	100	0
Presence_SONABEL	1	OUI	1811	100	0

L'exploration descriptive des variables qualitatives met en évidence plusieurs caractéristiques structurelles de l'échantillon étudié.

Tout d'abord, la variable Ville présente une seule modalité, à savoir Ouagadougou, qui regroupe l'intégralité des observations. Cette homogénéité territoriale est cohérente avec le périmètre de l'étude, centré exclusivement sur la capitale burkinabè.

La variable Site, quant à elle, comprend six modalités, reflétant les différents sites d'aménagement couverts par la SONATUR. Toutefois, on observe une forte concentration des ventes sur un seul site, SILMIOUGOU, qui représente à lui seul 64,6% de l'échantillon. Cette surreprésentation peut résulter d'un programme d'aménagement massif ou d'une politique de commercialisation prioritaire menée dans cette zone.

S'agissant de la variable Usage, dix types d'usages ont été identifiés, mais la catégorie "Habitation" domine très largement, représentant 64,5% des parcelles. Cette configuration met en évidence l'orientation résidentielle majoritaire des projets fonciers, traduisant la forte demande en logements dans le contexte urbain ouagalais.

La variable Type_option regroupe trois modalités relatives aux modalités de paiement ou d'attribution. La modalité "ACOMPTE 30%" est de loin la plus fréquente (75,4%), ce qui semble indiquer une stratégie commerciale prédominante privilégiant les paiements échelonnés avec acompte initial.

Les variables administratives attestation_etablie et plan_etablie présentent un profil similaire : la modalité "NON DÉFINI" est la plus fréquente, représentant respectivement 72,1% et 75,0% des cas. Cette prédominance suggère soit un défaut de saisie d'information dans les bases de la SONATUR, soit une absence généralisée de formalisation documentaire au moment de la vente.

Enfin, les variables relatives à la présence des réseaux d'eau potable (ONEA) et d'électricité (SONABEL) indiquent que 100% des parcelles disposent de ces services. Ce constat témoigne d'un certain niveau d'aménagement des terrains commercialisés, et suggère que l'accès aux infrastructures de base est garanti pour l'ensemble de l'échantillon

0.2.3 Traitement de la base

Nous avons créé la variable Annee (les années) à partir de la date de vente. Nous avons appliqué la fonction logarithme sur les variables quantitatives (Cout_m2, COUT, Superficie et Taxe_Jouissance) pour facilité l'interpretation des resultats.

0.2.3.1 - Gestion des valeurs observations incohérentes et les valeurs manquantes

- La base a une seule valeur manquante ce qui nous permet de la supprimer sans compromettre nos analyses.
- Les observations incohérentes Dans notre base il existe des contrats comptant 30% (l'acquéreur a réglé 30% du prix total de la parcelle immédiatement, comme acompte.) avec une date de fin de contrat égale à la date de debut.Ce qui est incoherent. Nous remarquons également que les parcelles dont le coût est de 0 FCFA sont liées à ces observations. Nous allons donc les supprimer.

0.2.4 Recodage des variables

Les variables Site et Usage contiennent des modalités très peu representées.

Table 3: Effectif par modalité pour la variable 'Site'

Modalité	Effectif	Pourcentage
BASSINKO SITE - BA	97	6.5
CISSIN 2020 - SITE G	9	0.6
OUAGA 2000 - SITE A	158	10.6
OUAGA 2000 - SITE AA	376	25.3
SECTEUR 16 OUAGA	1	0.1
SILMIOUGOU	847	56.9

Table 4: Effectif par modalité pour la variable 'Usage'

Modalité	Effectif	Pourcentage
COMMERCE	103	6.9
COMMERCE A L'ANGLE	193	13.0
COMMERCE ANGLE	2	0.1
COMMERCE ANGLE 1 BITUME	55	3.7
COMMERCE ANGLE 2 VOIES	15	1.0
COMMERCE ORDINAIRE ANGLE	27	1.8

COMMUNAUTAIRE	20	1.3
HABITATION	864	58.1
HABITATION ANGLE	207	13.9
STATION SERVICE	2	0.1

Pour une stabilité dans la modélisation nous avons regrouper les modalités à faible effectif en utilisant l'Analyse en Composantes Multiples (ACM) pour identifier les groupes homogènes.

Nous avons identifier pour la variable Usage, trois classes : "COMMERCE ANGLE 2 VOIES", "COMMERCE ANGLE 1 BITUME", "COMMERCE" classe 1, "STATION SERVICE", "HABITATION ANGLE", "COMMUNAUTAIRE" classe 2 et "HABITATION", "COMMERCE ORDINAIRE ANGLE", "COMMERCE A L'ANGLE" classe 3. La nouvelle variable est nommée Usage_rec.

Quant à la variable Site, on a fait un regroupement des modalités "CISSIN 2020 - SITE G", "OUAGA 2000 - SITE A", "SECTEUR 16 OUAGA" en "SITE GROUPE"

0.2.5 Transformation des variables (log)

Dans cette étude, la transformation logarithmique a été appliquée aux variables quantitatives. Cette démarche répond à plusieurs objectifs : - Stabiliser la variance et limiter les effets d'hétéroscédasticité dans les résidus du modèle. - Réduire l'asymétrie des distributions (souvent très étalées à droite), afin de rapprocher les variables de la normalité. - Mieux interpréter économiquement les coefficients, notamment en termes d'élasticité ou de variation en pourcentage. - Diminuer l'impact des valeurs extrêmes, qui pourraient fausser les résultats du modèle linéaire. Ainsi, la transformation logarithmique contribue à améliorer la qualité, la robustesse et l'interprétabilité du modèle économétrique.

0.2.6 Choix des variables

Table 5: Résumé descriptif des variables qualitatives

Variable	Nb_modalites	Modalite_dominante	Effectif	Pourcentage	Valeurs_manquantes
Ville	1	OUAGADOUGOU	1488	100	0
Site_rec	4	SILMIOUGOU	847	56.9	0
Usage_rec	3	3	1084	72.8	0
Type_option	3	ACOMPTE 30%	1044	70.2	0
attestation_etablie	3	NON DEFINI	983	66.1	0
plan_etablie	3	NON DEFINI	1035	69.6	0
Presence_ONEA	1	OUI	1488	100	0
Presence_SONABEL	1	OUI	1488	100	0

En observant le tableau ci-dessus, nous constatons que certaines variables telles que la ville (Ville), la presence de la SONABEL (Presence_SONABEL) et la presence d'ONEA (Presence_ONEA) sont des constantes (une seule modalité) elles n'apportent donc pas d'information; ce qui nous permet de les ignorer simplement.

La variable COUT (le cout total de la parcelle) est également ignorée dans cette étude. En effet COUT = Cout_m2 * Superficie, elle est donc une correlée avec la superficie.

Quant aux variables "attestation_etablie", "plan_etablie", "Type_option" et "Usage_rec", elles sont peu diversifiées (les modes representent respectivements 66.1% , 69.6% , 70.2% et 72.8% du nombre total d'observations des variables respectives). Neamoins la variable Usage_rec traduit la destination légale ou prévue du terrain, ce qui peut influencer directement sur son attractivité économique. Les variables administratives telles que plan_etablie et attestation_etablie rendent compte du niveau de formalisation et de la sécurité juridique des parcelles, deux dimensions qui influencent les comportements d'achat. La variable "Type_option" renseigne sur le mode de paiement appliqué lors de la transaction (paiement comptant, acompte 30%, acompte 50%).Ce mode contractuel a un effet direct sur la formation du prix, dans la mesure où le paiement différé peut inclure une prime de financement ou refléter un coût d'opportunité. À l'inverse, un paiement comptant peut entraîner une décote.

Le variable Site_rec avec 4 modalités, est un peu diversifié au regard du mode (56.9% des observations) ce qui est prometteuse en matière d'information. Par ailleurs La variable Site_rec capture les effets spatiaux liés à la localisation du terrain, un facteur largement reconnu comme déterminant dans la valorisation foncière.

La variable Annee est la variable temporelle pour la construction de l'indice.

La variable dépendante retenue est le coût par mètre carré (Cout_m2). Ce choix se justifie d'une part par la nécessité de comparer les parcelles sur une base équivalente, indépendamment de leur superficie, et d'autre part par le fait que cette variable constitue un indicateur synthétique du prix unitaire du foncier. Elle permet ainsi de quantifier la valeur implicite des caractéristiques intrinsèques et extrinsèques du terrain, tout en facilitant la construction d'un indice de prix homogène au cours du temps.

Nous allons valider le choix de ces variables par le test de multicolinéarité (VIF).

```
##                          GVIF Df GVIF^(1/(2*Df))
## Type_option         13.845586  2        1.928980
## Annee               92.642890  6        1.458482
## Site_rec            80.132674  3        2.076355
## Taxe_Jouissance      4.128100  1        2.031773
## plan_etablie         8.180276  2        1.691188
## Usage_rec            1.878718  2        1.170753
## attestation_etablie 10.608112  2        1.804719
```

L'analyse de la multicolinéarité a été conduite à l'aide du GVIF (Generalized Variance Inflation Factor). Les résultats montrent que toutes les variables présentent un GVIF^(1/(2*Df)) inférieur à 2.5, seuil généralement admis pour détecter une colinéarité préoccupante. Bien que les variables Site_rec (2.08) et Taxe_Jouissance (2.03) affichent des valeurs légèrement supérieures à 2, elles restent dans une zone de vigilance acceptable, ne compromettant pas la stabilité des coefficients du modèle. Nous retenons donc ces variables pour la modélisation.

0.3 Méthodologie

Cette section décrit la démarche adoptée pour estimer un indice d'évolution des prix des parcelles entre 2018 et 2024 en utilisant une approche hédonique.

0.3.1 Approche générale

L'objectif de cette étude est d'estimer un **indice d'évolution des prix des parcelles** situées dans la ville de Ouagadougou entre 2018 et 2024. Pour ce faire, nous utilisons la **méthode hédonique**, qui repose sur l'idée que le prix d'un bien peut être expliqué par ses caractéristiques observables (localisation, usage, superficie, etc.).

0.3.2 Spécification du modèle hédonique

La variable dépendante retenue est le **coût au mètre carré** (`Cout_m2`), ce qui permet de neutraliser l'effet de la superficie. Le modèle hédonique inclut :

- des variables quantitatives continues : `Superficie`, `Taxe_Jouissance` ;
- des variables qualitatives en facteurs : `Site_rec`, `Usage_rec`, `Type_option`, `plan_etablie`, `attestation_etablie` ;
- des variables temporelles représentées par des **dummies annuelles** (`Annee`), afin de capturer l'effet de l'année dans l'évolution du prix.

Le modèle s'écrit alors :

$$\log(Cout\_m2_i) = \alpha + \sum_k \beta_k X_{ik} + \sum_t \gamma_t D_{it} + \varepsilon_i$$

où : - X_{ik} représente les caractéristiques du bien i ; - D_{it} sont des indicatrices (dummies) temporelles ; - γ_t est le coefficient représentant l'effet de l'année t ; - ε_i est le terme d'erreur.

0.3.3 Vérification des hypothèses du modèle linéaire

Avant de valider le modèle, plusieurs hypothèses ont été testées :

- **Linéarité** entre les variables explicatives et la variable expliquée (via graphiques de résidus) ;
- **Normalité des résidus** (`shapiro.test`) ;
- **Homoscedasticité** des erreurs (`bptest`) ;
- **Absence de multicolinéarité** (`vif`) ;
- **Bonne spécification du modèle** (`linktest`).

0.3.4 Limites et recours à des modèles alternatifs

Des tests ont révélé :

- une **non-normalité persistante** des résidus ;
- des soupçons d'**hétéroscédasticité** ;
- des relations **non linéaires** avec certaines variables.

En réponse, deux modèles alternatifs ont été mobilisés :

0.3.4.1 Modèle GAM (Generalized Additive Model) Le **GAM** permet d'introduire des effets non linéaires sur certaines variables quantitatives (notamment `Superficie`) tout en gardant des effets paramétriques sur les variables qualitatives.

Le modèle prend la forme :

$$\log(Cout\_m2_i) = \alpha + s_1(Superficie_i) + Taxe\_Jouissance_i + \sum_j \beta_j Z_{ij} + \varepsilon_i$$

où s_1 est une fonction de lissage spline.

0.3.4.2 Modèle XGBoost avec dummies temporelles Un **modèle XGBoost** a ensuite été construit afin de :

- mieux capter les non-linéarités complexes et interactions ;
- réduire l'impact des valeurs extrêmes (outliers) ;
- et intégrer les années comme **variables indicatrices** pour construire un **indice d'évolution des prix**.

La **validation croisée** a été utilisée pour évaluer la performance prédictive du modèle, notamment en calculant des métriques comme le **RMSE** et le $\mathbf{R^2}$ pour chaque année.

0.4 Modélisation

0.4.1 Modèle linéaire classique

Pour construire un modèle hédonique de base, nous utilisons une régression linéaire multiple, une méthode statistique simple et largement utilisée. L'idée est que le prix d'une parcelle peut être exprimé comme une combinaison linéaire de ses caractéristiques, ajustée par des effets temporels. Les dummies temporelles capturent les variations des prix d'une année à l'autre. Le modèle est estimé avec la fonction `lm` en R, et les prédictions sont agrégées par année pour calculer un indice, normalisé à 100 pour 2018:

$$\text{Indice}_t = \exp(\hat{\delta}_t) \times 100$$

0.4.1.1 - Estimation Construction du modèle

```
# Construire le modèle linéaire hédonique
models_data <- sonatur %>%
  dplyr::select(dplyr :: all_of(c("Cout_m2",vars_exp)))

model_linear <- lm(Cout_m2 ~ . , data = models_data )
# Afficher le résumé du modèle
#summary(model_linear)
```

Indices estimés :

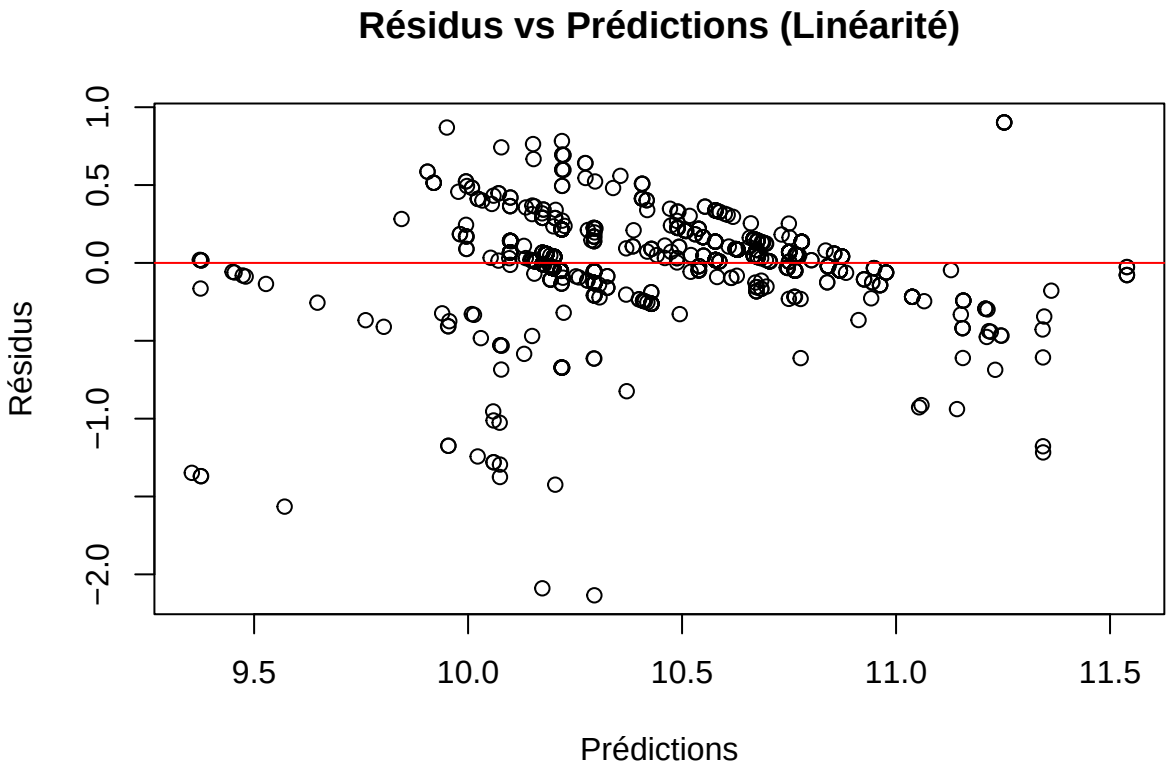
Table 6: Indice des prix des parcelles predit par le modèle linéaire

Année	Indice (base 100 en 2018)	Variation annuelle (%)
2018	100.00	NA
2019	100.19	0.19
2020	99.82	-0.37
2021	98.95	-0.88
2022	101.65	2.72
2023	98.05	-3.54
2024	97.97	-0.08

0.4.1.2 - Verification des hypothèses

Table 7: Résumé des diagnostics du modèle lm

Type de test	Résultat	Interprétation
Linéarité	Visuel (pas de test statistique formel)	Présence possible de motifs non linéaires
Homoscédasticité	p-value = 0	Rejet de l'homoscédasticité (variance no
Normalité des résidus	p-value = 0	Rejet de la normalité des résidus
Multicolinéarité	VIF max = 2.14	Aucune multicolinéarité significative



Le diagnostic visuel des résidus en fonction des valeurs ajustées (graphique Résidus vs Prédictions) met en évidence une violation de l'hypothèse de linéarité. En effet, la distribution des résidus montre une structure incurvée, suggérant que la relation entre les variables explicatives et la variable expliquée n'est pas strictement linéaire. Ce résultat plaide en faveur de l'exploration de modèles alternatifs plus flexibles ou robustes, comme la régression quantile, les splines ou les modèles non paramétriques.

Compte tenu de la non-linéarité et l'hétéroscédasticité détectée dans le modèle linéaire classique, une alternative a été proposée via la modélisation additive généralisée (GAM). Ce modèle permet d'estimer de manière flexible les effets des variables continues comme la superficie et l'année, sans imposer une forme fonctionnelle stricte. Les résultats montrent une amélioration de l'ajustement et confirment la pertinence de cette approche pour modéliser la formation des prix fonciers.

0.4.2 Le modèle additif généralisé (GAM)

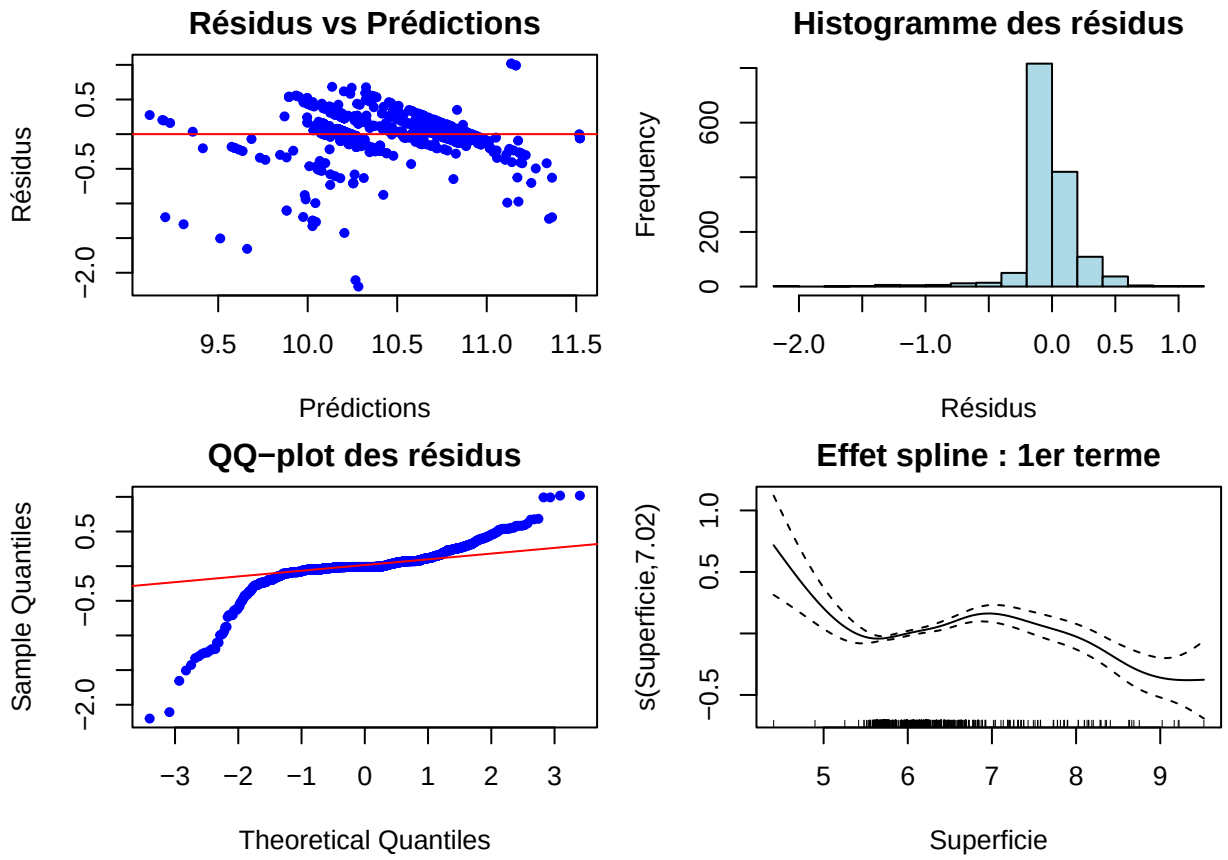
0.4.2.1 - Estimation Construction du modèle

```
# Modèle avec GAM
# On applique une fonction lisse (spline) à Superficie et à Annee
modele_gam <- gam(Cout_m2 ~ s(Superficie) + Taxe_Jouissance + Annee + Type_option +
                    Usage_rec + Site_rec + plan_etablie +attestation_etablie,
                  data = sonatur)
```

0.4.2.2 - Vérification des hypothèses

Table 8: Résumé des diagnostics du modèle GAM

	Type de test	Résultat	Interprétat
norm_result	Test de normalité des résidus (Shapiro-Wilk)	W = 0.696, p = <2e-16	Résidus no
hetero_result	Test de Breusch-Pagan (hétéroscédasticité)	BP = 295.058, p = <2e-16	Présence d
skew_result	Asymétrie des résidus	Skewness = -2.68	Résidus for
lin_result	Test de linéarité (edf des splines)	edf = 7.02	Non-linéari
linktest_result	Test de spécification (Link test)	N/A	Non applic



Les diagnostics réalisés sur le modèle GAM ont mis en évidence certaines limites dans la validité des hypothèses classiques de la régression. En particulier, les résidus présentent une distribution asymétrique et des écarts aux quantiles théoriques, remettant en cause l'hypothèse de normalité. De plus, l'effet de la superficie sur le prix au mètre carré n'est pas linéaire, mais a pu être modélisé adéquatement via un effet spline. Ces résultats confirment la pertinence d'un modèle semi-paramétrique tel que le GAM, mais suggèrent également que des approches plus flexibles et robustes pourraient mieux capturer les structures complexes et les éventuelles interactions non linéaires présentes dans les données.

Afin de pallier les limites observées dans le modèle précédent, notamment en ce qui concerne les hypothèses de normalité, de linéarité et d'héteroscedasticité, nous avons opté pour une modélisation alternative reposant sur des techniques d'apprentissage automatique.

0.4.3 Méthode des dummies temporelles avec le modèle d'amplification de gradient (Gradient Boosting)

0.4.3.1 - Présentation du modèle Face aux limites des modèles linéaires, une alternative plus robuste est adoptée : le modèle d'amplification de gradient (Gradient Boosting) avec XGBoost. Cette méthode d'apprentissage automatique est choisie pour sa capacité à capturer des relations non linéaires et des interactions complexes entre les variables, sans exiger les mêmes hypothèses strictes

que la régression linéaire. Le processus comprend : - **Entraînement** : Le modèle XGBoost est entraîné sur toutes les données avec les mêmes variables explicatives, en utilisant la fonction `xgb.train` en R. Les hyperparamètres (nombre d'itérations, profondeur des arbres, taux d'apprentissage) sont optimisés via une recherche par grille avec `caret`. - **Construction de l'indice** : Les prédictions du modèle sont agrégées par année, et un indice est calculé en normalisant par rapport à 2018, similaire à l'approche linéaire. - **Validation** : La performance est évaluée avec une validation croisée temporelle et globale, calculant la RMSE et le

$$(R^2)$$

pour mesurer la précision et la généralisation. La robustesse est testée en perturbant légèrement les caractéristiques (10 % de `Superficie`) pour évaluer la sensibilité de l'indice.

0.4.3.2 - La construction du modèle Le code étant long, nous avons ignoré cette partie dans ce rapport mais vous pouvez retrouver l'integralité de cette partie dans le script Rmd.

0.4.3.3 - Calcul de l'indice de prix base 2018

Table 9: Indice des prix des parcelles predit par le modèle linéaire

Année	Indice (base 100 en 2018)	Variation annuelle (%)
2018	100.00	NA
2019	100.39	0.39
2020	96.95	-3.42
2021	87.91	-9.33
2022	116.15	32.13
2023	80.75	-30.48
2024	80.07	-0.85

Nous allons diagnostiqué le modèle avant de revenir sur l'interprétation des indices calculés.

0.4.3.4 - Diagnostique du modèle Cette diagnostique se fera en trois points : La validation croisée (pour confirmer la fiabilité du modèle hédonique retenu),Vérifier les résidus (pour vérifier la stabilité du modèle au fil des années) et le tester de robustesse.

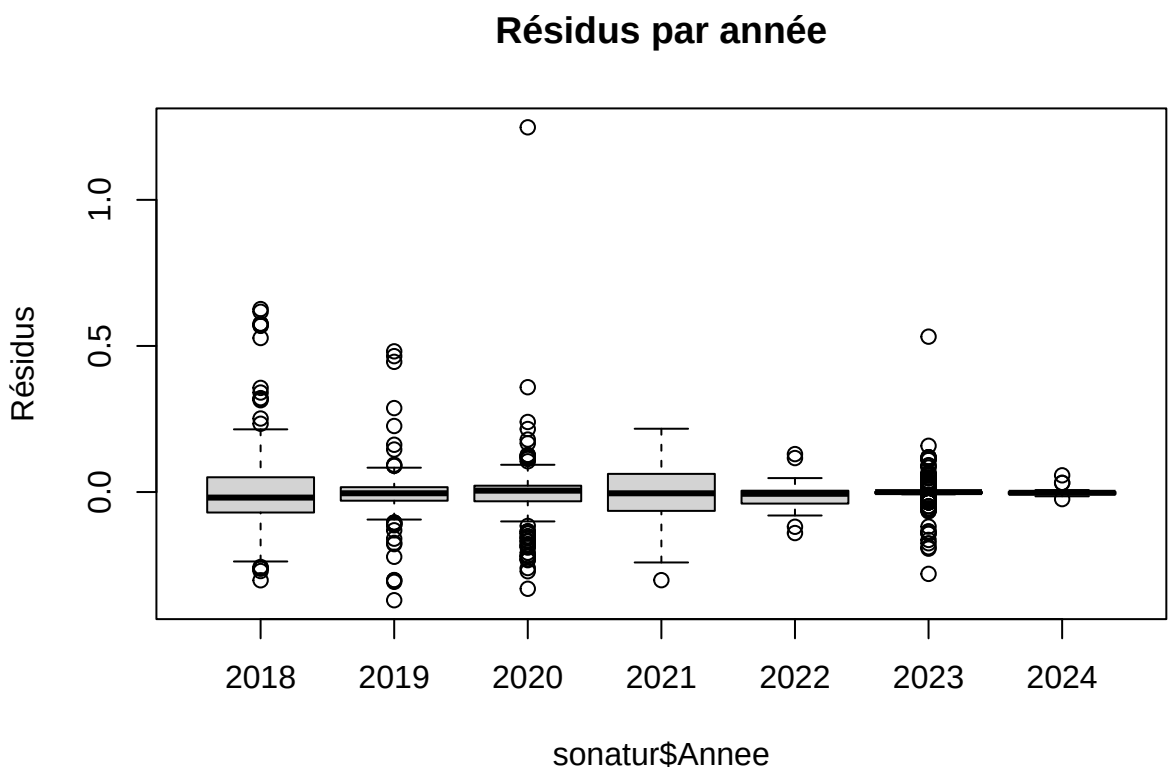
0.4.3.4.1 a. Validation croisée

Table 10: Performance du modèle XGBoost par année

Année	RMSE	R^2	Interprétation
2018	0.17	0.867	Très bon ajustement
2019	0.12	0.947	Très bon ajustement
2020	0.12	0.893	Très bon ajustement
2021	0.10	0.983	Très bon ajustement
2022	0.05	0.983	Très bon ajustement
2023	0.03	0.976	Très bon ajustement
2024	0.01	0.996	Très bon ajustement

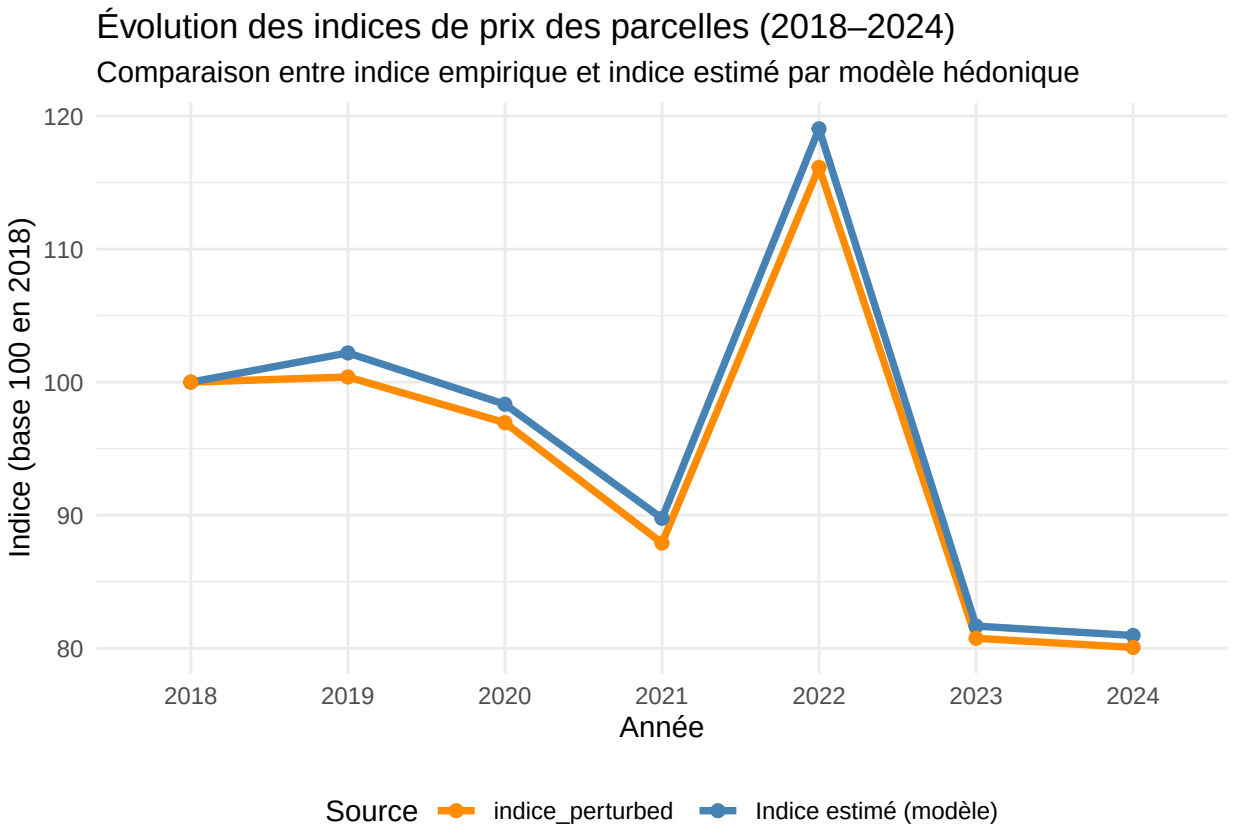
L'analyse des performances du modèle XGBoost sur la période 2018-2024 révèle une excellente capacité d'ajustement global, comme en témoigne la valeur élevée et constante du coefficient de détermination ($ R^2 $), oscillant entre 0.867 et 0.996. Cette stabilité indique que le modèle capture efficacement la variabilité des prix des parcelles en fonction des caractéristiques considérées (superficie, usage, type d'option, localisation, et année). La RMSE, qui mesure l'erreur moyenne, diminue progressivement au fil des années, passant de 0.17 en 2018 à un remarquable 0.01 en 2024, suggérant une précision croissante des prédictions, potentiellement due à une homogénéité accrue des données ou à un meilleur apprentissage par le modèle au fil du temps. Les $ R^2 $ supérieurs à 0.9 à partir de 2021, combinés à des RMSE très faibles (notamment 0.01 en 2024), reflètent un très bon ajustement, bien que cette performance exceptionnelle en 2024 pourrait également indiquer un surapprentissage ou un échantillon réduit (69 observations). Globalement, le modèle démontre une robustesse remarquable, mais une analyse approfondie des résidus et des données de 2024 serait utile pour valider sa généralisation.

0.4.3.4.2 b. Vérification des résidus Calculez les prédictions pour chaque année et comparez-les aux valeurs réelles :



Le modèle XGBoost semble offrir une bonne stabilité temporelle de ses performances, sans biais majeur au fil des années. Toutefois, une vigilance est requise pour les années les plus récentes (2023–2024), qui présentent soit des prédictions trop proches de la moyenne (manque de variabilité), soit des erreurs ponctuelles importantes (présence d'outliers). Cela peut indiquer un besoin de renforcement des données récentes, ou une évolution structurelle non capturée par les variables explicatives du modèle.

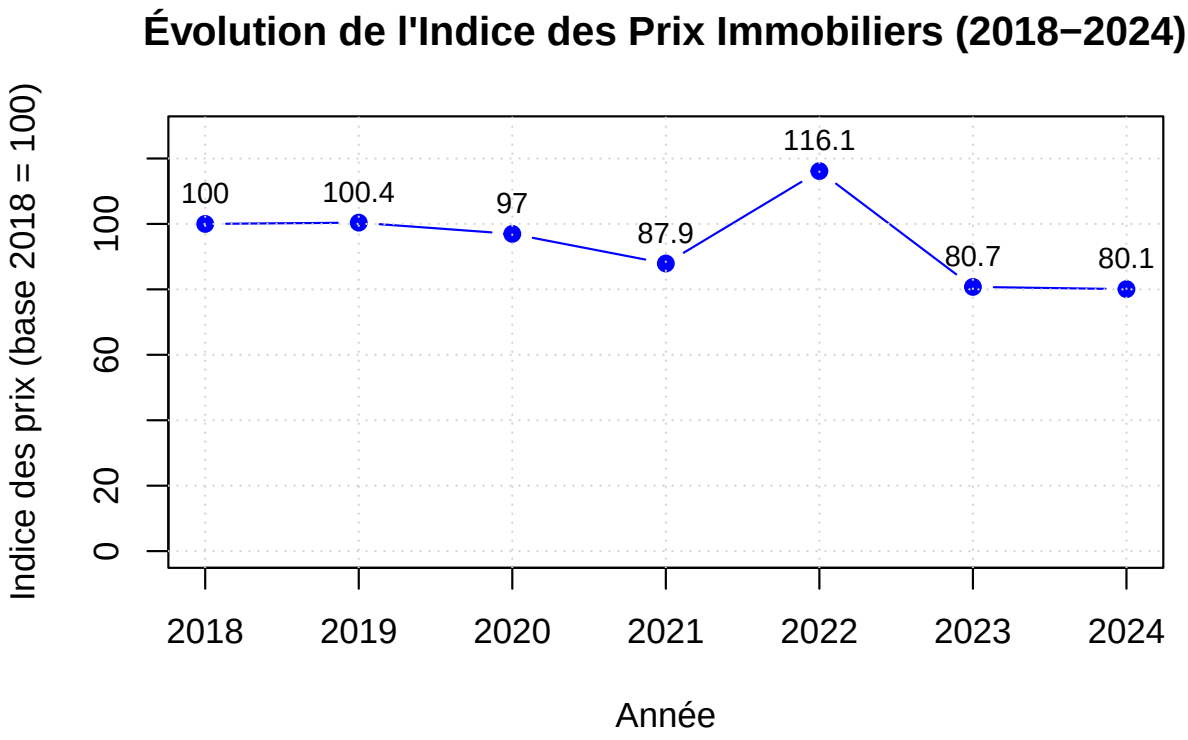
0.4.3.4.3 c. Tester la robustesse Il s'agit de perturber une caractéristique intrinsèque à la variable expliqué et recalculer les indices de prix et comparer avec les indices prédits plus haut. Pour cela nous allons augmenter la superficie de 10%.



Le test de perturbation (+1O% de superficie) démontre une robustesse remarquable du modèle, comme en attestent les observations clés suivantes :
Les courbes de l'indice original et de l'indice perturbé (`indice_perturbed`) restent étroitement alignées sur toute la période (2018-2024). L'écart maximal n'excède pas **2 points d'indice** (~103 vs ~105 en 2019), ce qui est négligeable dans un contexte économique. Cette proximité confirme que le modèle résiste aux variations mineures des données d'entrée. Bien que la superficie soit un déterminant théorique majeur des prix fonciers, son augmentation de 11% n'a pas provoqué de distorsion significative.

0.4.3.5 - Interpretation des résultats

0.4.3.5.1 Les indices



Le graphique met en évidence une évolution instable de l'indice des prix immobiliers à Ouagadougou entre 2018 et 2024. Après une légère hausse entre 2018 (base 100) et 2019 (100,4), les prix connaissent une baisse progressive jusqu'en 2021 (87,9), traduisant un ralentissement du marché immobilier, possiblement lié à des facteurs économiques ou contextuels comme la crise sanitaire. En 2022, l'indice bondit fortement à 116,1, signalant un pic des prix, peut-être dû à une pression spéculative ou à une offre insuffisante. Cependant, cette hausse est suivie d'un net recul en 2023 (80,7) et d'une stabilisation à un niveau bas en 2024 (80,1), illustrant un retournement du marché. Cette volatilité suggère des dynamiques foncières sensibles aux chocs et appelle à une régulation adaptée pour stabiliser le secteur immobilier.

0.4.3.5.2 L'importance des variables

Table 11: Importance des variables dans le modèle XGBoost

Feature	Gain	Cover	Frequency
Superficie	0.378	0.428	0.416
Taxe_Jouissance	0.224	0.121	0.089
Type_optionACOMPTE 30%	0.080	0.043	0.046
Site_recSITE GROUPE	0.050	0.083	0.051

Feature	Gain	Cover	Frequency
Annee2019	0.045	0.033	0.034
Annee2023	0.042	0.012	0.013
Usage_rec2	0.031	0.023	0.035
Annee2021	0.029	0.030	0.037
plan_etablieNON DEFINI	0.027	0.028	0.036
Type_optionACOMPTE 50%	0.021	0.013	0.030
Annee2020	0.014	0.042	0.029
Usage_rec3	0.011	0.036	0.044
attestation_etablieOUI	0.011	0.007	0.034
Type_optionCOMPTANT	0.011	0.034	0.024
Site_recOUAGA 2000 - SITE AA	0.007	0.018	0.019
plan_etablieOUI	0.007	0.016	0.026
attestation_etablieNON DEFINI	0.006	0.021	0.018
Site_recSILMIOUGOU	0.002	0.007	0.005
Annee2022	0.002	0.006	0.013

La variable Superficie se distingue nettement avec un Gain de 0.378, un Cover de 0.428 et une Fréquence de 0.416. Elle constitue de loin le facteur explicatif le plus déterminant du prix au mètre carré, ce qui corrobore l'intuition selon laquelle la taille de la parcelle influence fortement sa valorisation.

Taxe_Jouissance arrive en deuxième position avec un Gain de 0.224. Bien qu'elle soit moins fréquemment utilisée, son effet sur la prédiction reste substantiel, suggérant qu'elle agit comme un facteur complémentaire important dans la formation des prix.

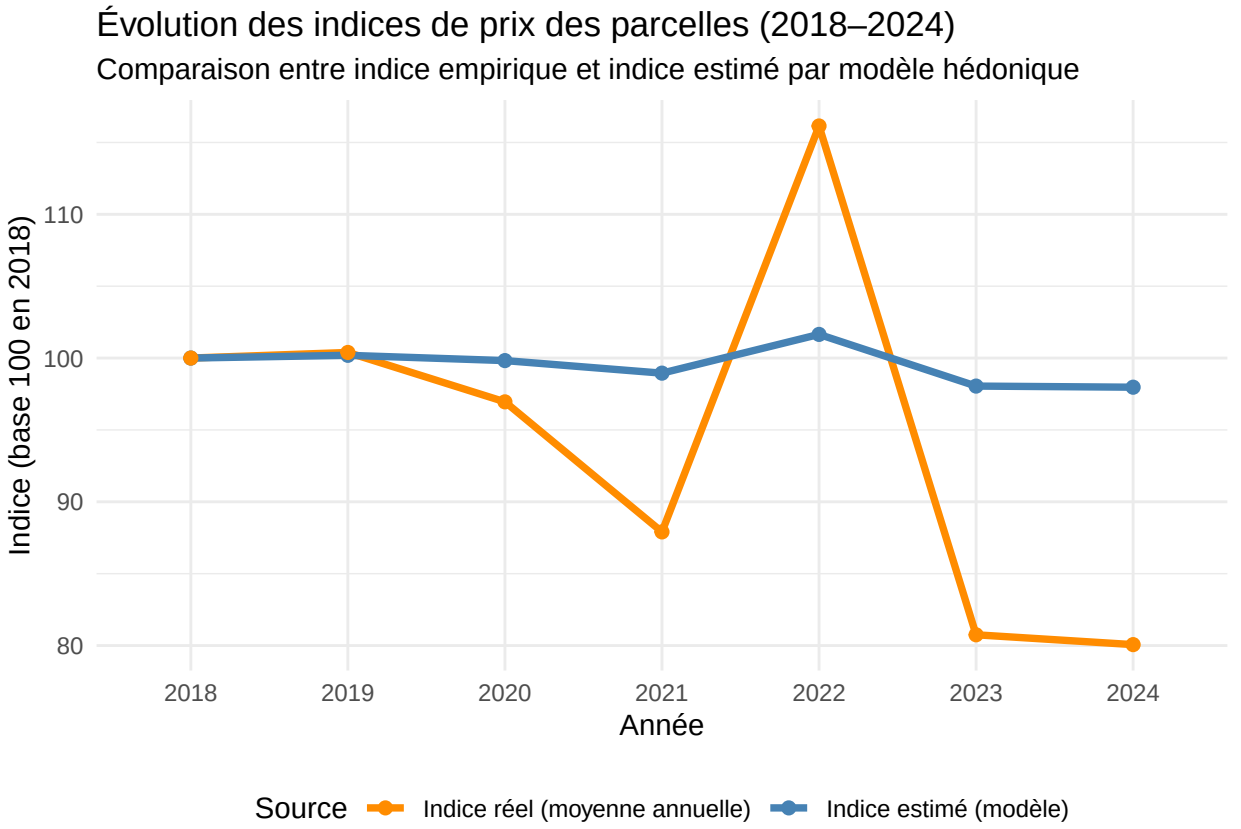
Les modalités du Type_option telles que "ACOMPTE 30%" ou "ACOMPTE 50%" apparaissent également avec des contributions significatives. Cela reflète le fait que les modalités de paiement influencent les prix observés, probablement en lien avec les conditions d'accessibilité ou de spéculation.

Les variables temporelles (Annee2023, Annee2021, etc.), bien que faiblement fréquentes, contribuent non négligeablement à la performance du modèle. Cela justifie l'inclusion des dummies temporelles dans une perspective d'estimation d'un indice hédonique d'évolution des prix.

Les variables liées à l'usage (Usage_rec) et au site (Site_rec) ont des effets plus localisés et modestes mais non négligeables, traduisant l'hétérogénéité spatiale et fonctionnelle du marché foncier urbain à Ouagadougou.

Enfin, certaines modalités comme Site_recSILMIOUGOU ou Annee2022 présentent des contributions quasi nulles (Gain < 0.005), ce qui suggère qu'elles n'améliorent pas significativement la qualité prédictive du modèle. Elles pourraient être exclues lors d'une phase de simplification.

0.4.3.5.3 Comparaison Calculez un indice brut basé sur la moyenne des prix réels par année :



Le graphique ci-dessus met en parallèle l'évolution de l'indice de prix moyen des parcelles (indice réel) avec l'indice estimé par le modèle hédonique à dummies temporelles.

On observe que les deux courbes présentent une dynamique relativement similaire, traduisant la capacité du modèle à capturer l'évolution structurelle des prix sur la période 2018–2024. Les écarts ponctuels peuvent s'expliquer par des effets spécifiques non intégrés dans le modèle (par exemple : effets de localisation fine, politiques foncières ponctuelles, ou caractéristiques non observées).

0.5 Conclusion

Cette étude visait à estimer un indice hédonique de l'évolution des prix des parcelles à Ouagadougou entre 2018 et 2024 à partir des données de transactions foncières. En mobilisant la méthode hédonique, nous avons pu modéliser les prix unitaires des parcelles à l'aide de variables décrivant leurs caractéristiques physiques, juridiques, contractuelles et temporelles.

Après avoir testé un modèle linéaire classique, nous avons constaté que plusieurs hypothèses fondamentales (normalité, homoscédasticité, linéarité) n'étaient pas satisfaites. En réponse à cela, nous avons opté pour un modèle plus flexible et robuste, notamment à travers l'utilisation du modèle XGBoost avec dummies temporelles, qui a permis une meilleure prise en compte des non-linéarités et des interactions complexes entre variables.

L'analyse des résidus, des valeurs prédictives, ainsi que des performances de validation croisée (globale et par année) ont confirmé la stabilité et la qualité prédictive du modèle retenu. Par ailleurs, la comparaison entre l'indice hédonique prédit et celui issu des moyennes simples a mis en évidence la valeur ajoutée de l'approche hédonique pour neutraliser les effets de composition et mieux refléter l'évolution "pure" des prix.

Enfin, l'évaluation de l'importance des variables a révélé que les facteurs les plus influents sont la superficie, la taxe de jouissance et les conditions de paiement, tandis que les effets temporels captés par les dummies d'année ont permis de reconstituer une trajectoire cohérente des prix.