

*Master Econométrie et Statistique Appliquée*

2024-2025

PROJET D'ÉTUDE MASTER 1 ESA (INSEE – LEO)

## Analyse économétrique des mutations immobilières et des salaires à un niveau territorial



*Auteurs*

LACROIX Ewan

KARAPETYAN Marieta

NOËL Julien

ROMAIN Canelle

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Postulats initiaux . . . . .	3
<b>2</b>	<b>Entrevue globale</b>	<b>5</b>
2.1	Echelle nationale . . . . .	5
2.2	Par départements . . . . .	8
2.3	Par zones d'emploi . . . . .	9
<b>3</b>	<b>Modèle Spatial</b>	<b>12</b>
3.1	Notions fondamentales . . . . .	12
3.1.1	L'autocorrelation spatiale . . . . .	12
3.1.2	Matrice de Pondération Spatiale . . . . .	12
3.1.3	L'indice de Moran . . . . .	14
3.2	Modèles SAR, SEM, SDM . . . . .	17
3.2.1	Choix du modèle . . . . .	17
3.2.2	Application et Résultats . . . . .	19
3.2.3	Interprétation Économique . . . . .	20
3.2.4	Variance Expliquée de notre modèle : Interprétation des effets spatiaux . . . . .	22
3.2.5	Analyse des Résidus du modèle SDM . . . . .	23
3.3	Modèle GWR . . . . .	25
3.3.1	Modèle simple . . . . .	25
3.3.2	Modèle enrichi . . . . .	27
3.3.3	Analyse des résidus . . . . .	28
3.4	Analyse complémentaire . . . . .	30
3.4.1	Transactions expliquées par les salaires . . . . .	30

3.4.2	Etudes sur années antérieures . . . . .	32
<b>4</b>	<b>Panel sur Zones d'emploi</b>	<b>34</b>
4.1	Analyse préliminaire des données . . . . .	34
4.1.1	Paramétrage de la structure spatiale . . . . .	35
4.1.2	Spécification du modèle . . . . .	36
4.2	Estimation . . . . .	38
<b>5</b>	<b>Dimension temporelle</b>	<b>41</b>
5.1	Analyse de la stationnarité des séries . . . . .	41
5.2	Analyse de l'auto-corrélation . . . . .	44
5.3	Cointégration et modélisation . . . . .	45
5.3.1	Détection d'une cointégration . . . . .	45
5.3.2	Modélisation VECM . . . . .	48
5.4	Vérification des résidus du modèle . . . . .	51
<b>6</b>	<b>Conclusion</b>	<b>54</b>
<b>7</b>	<b>Bibliographie</b>	<b>55</b>
<b>8</b>	<b>Annexe</b>	<b>57</b>
8.1	Dimension temporelle . . . . .	58
8.2	Modèle Spatial . . . . .	58
8.3	Panel . . . . .	62
8.3.1	Variables . . . . .	62
8.3.2	L'impact des mutations sur les salaires . . . . .	65
8.3.3	Sous annexe Panel . . . . .	79

## 1 Introduction

Dans un contexte économique de fortes disparités salariales, explorer le lien entre les mutations immobilières et les niveaux des salaires est une manière de comprendre les dynamiques locales qui façonnent nos territoires. Ce champ d'étude repose sur la conception d'une dynamique commune pour les territoires, plus précisément pour les zones dont les changements de propriétés se font de manière fréquente. Ce changement de propriétaire peut être l'œuvre d'une attractivité économique croissante, d'une mobilité affluente des résidents voire même d'une restructuration des liens socio-économiques. Au sein de notre projet conduit avec l'INSEE, l'objet de notre étude vise à analyser les relations entre le nombre de logements considérés par des changements de propriétaires et le niveau des salaires observé dans les zones d'emploi de la France.

Cette étude pose le cadre d'un indicateur économique spécifique pour caractériser le marché immobilier : le nombre de logements ayant changé de propriétaire. Cette analyse permet de capter l'intensité de la circulation essentielle sur un territoire au-delà des simples ventes de logements. Ce sujet nous est apparu comme une opportunité de croiser les mutations immobilières et les salaires afin de proposer une analyse des dynamiques socio-économiques.

L'idée fondamentale repose sur le fait que les mutations immobilières peuvent refléter une dynamique économique locale. Ces dynamismes peuvent être sous la forme de renforcement de l'attractivité des résidences ou même d'une allocation des ressources productives. À travers la mobilité des résidences qui surviennent, les mutations peuvent être révélatrices de phénomènes dits spatiaux ou même de sélection de capital humain, pour laquelle des travailleurs plus qualifiés migrent vers les zones les plus dynamiques (Glaeser and Gyourko,2005).

Notre projet sera constitué de plusieurs théories qui appuient le lien entre le marché de l'immobilier et les salaires locaux (Charruau and Epaillard,2017). Ces références permettent d'ancrer notre réflexion dans un cadre reconnu dans la littérature économique. Parmi celles-ci, nous pouvons citer l'article « Urban Decline and Durable Housing » de Glaeser et Gyourko montrant que les facteurs des prix immobiliers constituent une part influente de l'attractivité urbaine.

Plus précisément les auteurs démontrent que pour les zones où les logements sont dévalorisés, la baisse des prix immobiliers attire une population alternative modifiant ainsi la structure économique et salariale des villes. Cette influence de l'attractivité urbaine structure les trajectoires des individus en matière d'emplois. De plus l'approche du capital spatial de Andersson et Klaesson démontre bien

que les salaires sont fortement affectés par la structure géographique du logement et la proximité des bassins d'activité. En outre ces travaux éclairent notre problématique sous différents angles.

Pour affiner notre analyse, il conviendra de justifier nos travaux d'un point de vue empirique prouvant que la structure des marchés immobiliers influence les inégalités salariales locales. Des travaux ont mis en évidence que la mobilité résidentielle peut générer des effets de relocalisation des ménages en fonction des opportunités économiques, ce qui pourrait accentuer les écarts salariaux entre les zones d'emploi (Doeringerand Piore,2020). Un territoire se voudra attractif là où les transactions sont nombreuses et concentrera des opportunités professionnelles et des investissements économiques qui seront susceptibles de faire augmenter les salaires.

C'est dans ce cadre que nous allons mobiliser des méthodologies d'économétrie spatiale, en particulier le modèle SDM afin de modéliser les salaires comme une fonction des mutations immobilières. Ce choix de méthodologie permet de tenir compte des interactions géographiques entre les zones d'emploi et de distinguer les effets locaux et voisins de ces déterminants. La complexité des interactions entre les dynamiques immobilières et les niveaux de salaires met en lumière les profondes interdépendances entre le marché du logement, le marché du travail et la géographie économique. Cette ambivalence souligne la nécessité d'une lecture territorialisée des politiques publiques. Encourager les mutations dans certaines zones pourrait stimuler l'activité et les salaires, à condition de maîtriser les effets d'exclusion ou de diffusion vers d'autres territoires. En définitive, ce travail montre que les mutations immobilières ne sont pas de simples évènements locaux, mais qu'elles sont au cœur des recompositions socio-économiques locales. Mieux comprendre leur lien avec les salaires permettrait de guider les stratégies d'aménagement du territoire, de lutter contre les inégalités territoriales et de pilotage des politiques de logement et d'emploi.

## 1.1 Postulats initiaux

Il est clair que les salaires ne sont pas uniquement déterminés par des facteurs individuels tels que le niveau de formation ou l'expérience professionnelle. Ils s'inscrivent dans un environnement territorial structuré par des dynamiques économiques, sociales et immobilières. Parmi ces dynamiques, les mutations immobilières, et en particulier le nombre de logements changeant de propriétaires, apparaissent comme un indicateur pertinent de l'attractivité et de la vitalité d'un territoire. Notre intuition initiale repose sur l'idée que dans les zones d'emploi où le marché immobilier est plus actif, c'est-à-dire là où les logements changent plus fréquemment de propriétaires, nous observons sou-

vent une plus grande mobilité résidentielle, une recomposition démographique et potentiellement une demande plus forte de main-d'œuvre qualifiée. Ces éléments peuvent contribuer à tirer les salaires moyens vers le haut. Une telle dynamique pourrait être liée à des effets d'agglomération, à l'arrivée de populations à revenu plus élevé ou encore à une modernisation du tissu économique local. Cependant, cette intuition mérite d'être nuancée. Un nombre élevé de mutations peut également résulter d'un marché spéculatif, d'une rotation de population contrainte, ou d'un phénomène de gentrification, qui n'entraîne pas nécessairement une augmentation des salaires pour l'ensemble des travailleurs. De plus, dans les zones périphériques ou rurales, une hausse des mutations peut refléter une pression foncière sans véritable dynamique salariale, voire une fuite des travailleurs vers d'autres bassins d'emploi mieux rémunérés. Dès lors, il est légitime de se demander :

Les mutations immobilières agissent-elles comme un moteur ou comme un symptôme des écarts de salaires entre territoires ?

Les effets observés sont-ils strictement locaux, ou se diffusent-ils spatialement vers les zones voisines ?

Comment ces dynamiques interagissent-elles avec d'autres variables structurelles comme le taux de chômage ou la densité d'emploi ?

Ce questionnement justifie le recours à une analyse économétrique formelle. Il ne s'agit pas uniquement d'identifier une corrélation globale entre mutations et salaires, mais d'étudier la manière dont cette relation varie dans le temps, et si des externalités territoriales positives ou négatives influencent les résultats. Enfin, cette réflexion s'inscrit dans un contexte plus large de mutations socio-économiques, où les inégalités territoriales de revenus tendent à se creuser, et où les politiques publiques s'attachent à rééquilibrer les dynamiques territoriales. Une analyse fine et localisée des liens entre marché immobilier et salaires est donc indispensable pour mieux comprendre les ressorts de l'attractivité, et pour concevoir des politiques d'aménagement et d'emploi adaptées à chaque territoire.

## 2 Entrevue globale

### 2.1 Echelle nationale

Tout au long de ce rapport, nous parlerons de *transactions* et de *mutations* pour faire référence au même indicateur : le nombre de logements changeant de propriétaire.

Afin de mieux comprendre la relation entre le nombre de transactions immobilières et les niveaux de salaires annuels, cette section présente une analyse descriptive de ces variables. L'objectif est de dresser un portrait global des données dont nous disposons avant d'effectuer des analyses plus approfondies.

Les statistiques descriptives permettent de mettre en évidence les tendances générales, les évolutions dans le temps ainsi que les éventuelles corrélations visuelles entre les variables.

Les variables analysées à l'échelle nationale sur une période de 10 années sont :

- Le nombre de transactions immobilières (changements de propriétaire), indicateur de l'activité du marché.
- Le salaire net horaire moyen, représentatif du pouvoir d'achat des ménages.

Table 1: Statistiques des données

salaires	transactions
Min. :14.28	Min. : 753733
1st Qu.:14.60	1st Qu.: 901832
Median :15.10	Median :1022206
Mean :15.36	Mean :1004945
3rd Qu.:16.19	3rd Qu.:1075662
Max. :17.02	Max. :1251198

Nous remarquons que les salaires sont regroupés entre 14.28 et 17.02 et que la médiane, de 15.10, suggère une légère asymétrie positive. De plus, 50% des salaires sont entre 14.60 et 16.19, ce qui n'est pas une grande dispersion.

Pour les transactions, nous remarquons une variabilité du nombre de logements ayant changé de propriétaires oscillant entre 753 000 et 1 025 000, mais une grande partie restant autour des 900 000/1 000 000.

Par des analyses statistiques plus approfondies présentes en annexe, nous pouvons identifier des différences de répartition des salaires par sexe, CSP et tranche d'âge.

Le type de personnes qui ont les plus hauts salaires horaires sont les hommes cadres (25.90€/heure en moyenne).

Nous remarquons que les hommes gagnent plus que les femmes (14.98€/heure contre 12.68€/heure en moyenne).

Il apparaît que le salaire horaire augmente avec l'âge (9.978€/heure pour les 18/25 ans, puis 13.81€/heure pour les 26/49 ans, puis 16.23€/heure pour les plus de 50 ans).

Toutes ces données sont très intéressantes pour avoir une idée des rémunérations et des écarts de salaires selon certaines catégories et certains critères.

Nous pouvons représenter nos deux séries (les salaires et le nombre de transactions) graphiquement, ce qui nous donnera une idée plus claire sur le comportement de celles-ci.

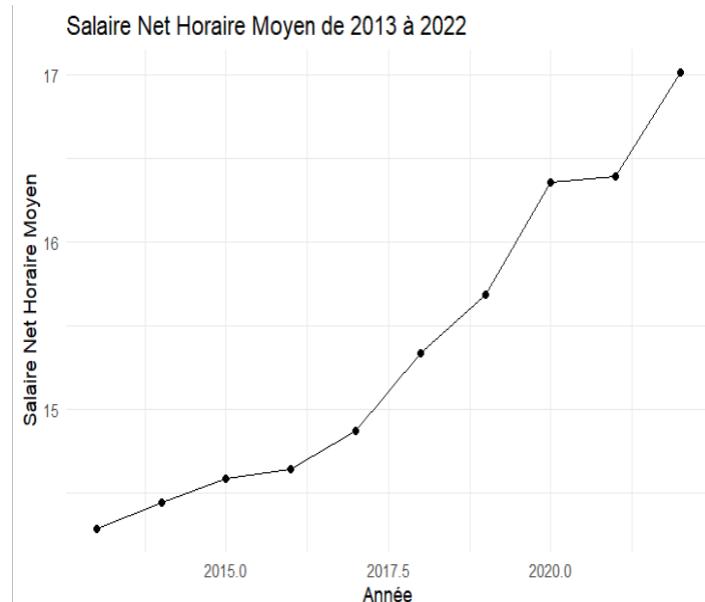


Figure 1: Courbe salaire

Nous constatons une tendance générale haussière des salaires horaires nets moyens sur la période. Toutefois, cette augmentation n'est pas linéaire et nous relevons une période de stagnation entre

2020 et 2021, coïncidant avec la période de ralentissement économique et d'incertitude causée par le Covid.

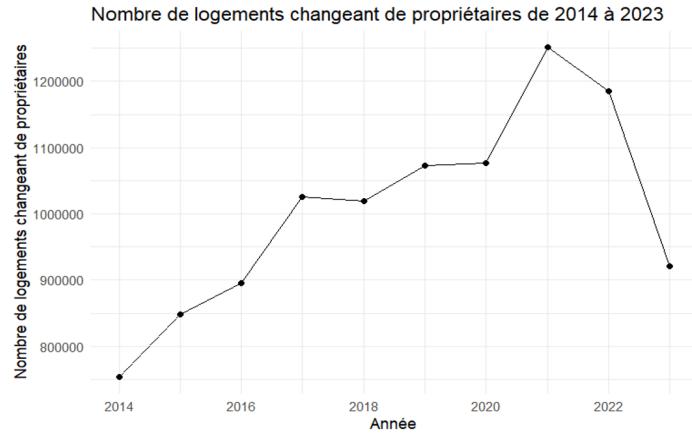


Figure 2: Courbe transactions

Depuis 2014, le nombre de logements changeant de propriétaire a toujours été en hausse. Cependant, nous observons qu'il a fortement baissé les dernières années pour atteindre près de 900 000 en 2023, tandis que 2 ans plus tôt, en 2021, ce chiffre s'élevait à plus d'1,2 million. Une des explications de ce changement soudain est certainement la crise du Covid qui a créé une stagnation en 2020 puis ce pic en 2021. En effet, l'année 2020 n'a pas été fructueuse pour réaliser des transactions immobilières, car les gens sont restés chez eux à cause du confinement et donc leur priorité n'était pas de changer de logement. Nous pouvons aussi penser à la hausse du taux directeur pour expliquer les chiffres bas de 2023.

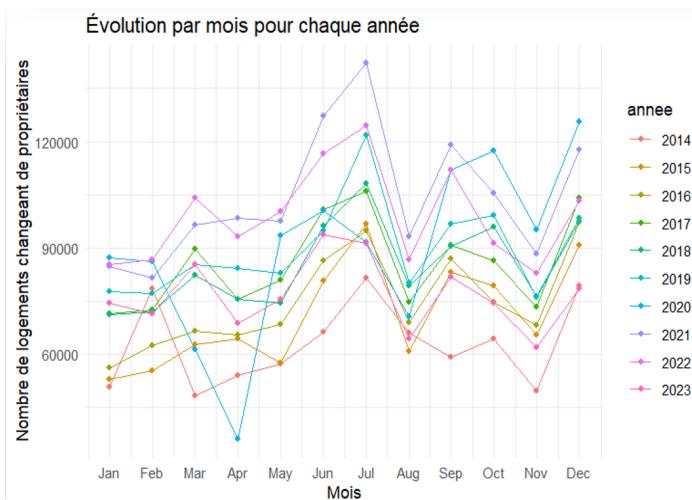


Figure 3: Courbes transactions

Chaque année, nous constatons une tendance d'évolution similaire. L'été semble être une période où il y a le plus de transactions, puis il y a une baisse marquée en août. Peut-être est-ce dû au fait que les administrations et les acheteurs sont en vacances durant ces périodes. Nous observons clairement un motif récurrent dans le marché immobilier. Nous remarquons une forte baisse en mars puis avril 2020 qui confirme que le marché de l'immobilier a fortement subi la crise pandémique. Globalement, le même schéma se répète chaque année sur le marché de l'immobilier.

## 2.2 Par départements

Nous décidons maintenant de choisir l'année 2022 pour représenter géographiquement nos deux variables à l'échelle des départements puis des zones d'emploi. Le schéma géographique que nous observons est sensiblement le même pour toutes les années dont nous disposons et 2022 est une année récente et représentative.

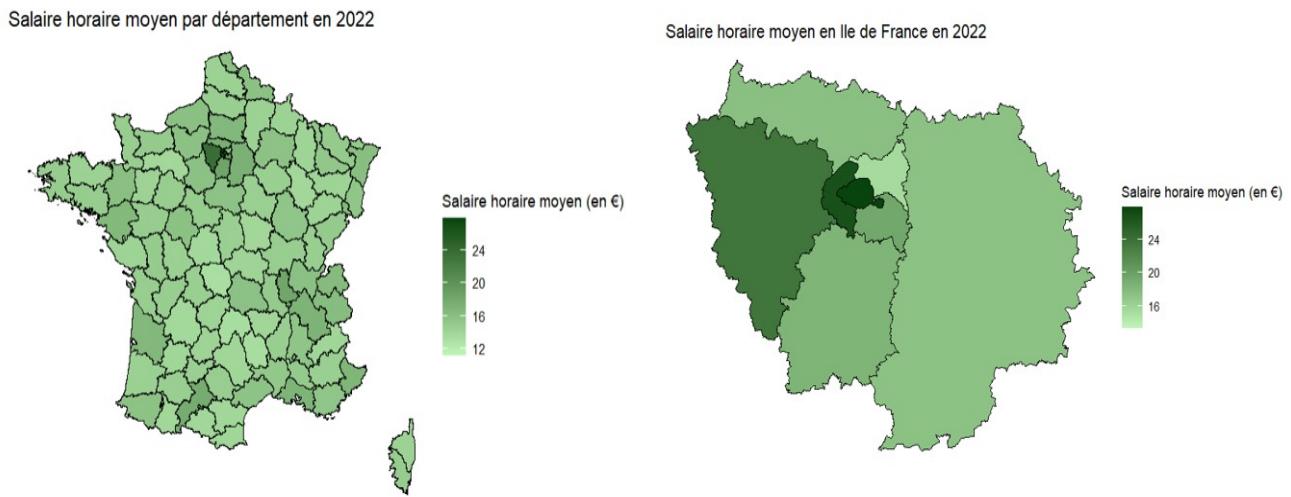


Figure 4: Carte salaire département

Sur cette carte, nous remarquons une homogénéité des salaires entre les départements. Cependant une zone se démarque en France : la région parisienne semble distinctement être un endroit où les salaires sont plus élevés que dans le reste de la France.

En effet, Paris, les Hauts-de-Seine et les Yvelines se démarquent par leurs salaires élevés. Cette zone offre de bonnes opportunités d'emploi, et accueille les sièges de la majorité des grandes entreprises, et accueille ainsi de nombreux cadres et professions mieux payées.

Nombre total de transactions immobilières par département en 2022

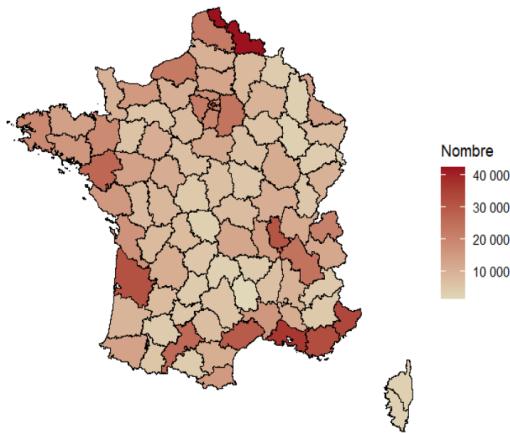


Figure 5: Carte transactions Departements

Nous remarquons une toute autre répartition du côté du nombre de transactions. En effet, les départements dans lesquels les plus grandes villes françaises se trouvent témoignent de plus de transactions immobilières. Cela s'explique par le dynamisme de leur bassin d'emploi et leur tissu économique diversifié. Les zones qui se démarquent le plus sont la Côte d'Azur (Alpes-Maritimes et Bouches-du-Rhône), avec son climat attractif et son fort tourisme, la Gironde, avec son économie dynamique et sa qualité de vie, Paris et ses départements alentours, avec sa forte densité urbaine et son attractivité internationale, le Nord, avec son marché immobilier dynamique, ce département étant un des plus peuplés de France.

### 2.3 Par zones d'emploi

La représentation par zones d'emploi peut être un critère plus intéressant pour notre analyse géographique. En effet, d'après l'Insee, *une zone d'emploi est un ensemble de communes dans lequel la plupart des actifs résident et travaillent, et où les établissements peuvent trouver l'essentiel de leur main-d'œuvre*. Ce découpage du territoire capte plus efficacement les dynamiques économiques locales.

Comme observé pour les départements, la région parisienne et ses alentours sont les lieux où les salaires nets horaires moyens sont les plus élevés en France. Les zones d'emploi de Versailles-Saint-Quentin, Paris, Seine-Yvelinoise, Rambouillet et Saclay se classent parmi celles proposant les plus hauts salaires, avec un salaire horaire net moyen supérieur à 20€/heure, bien supérieur à la moyenne.

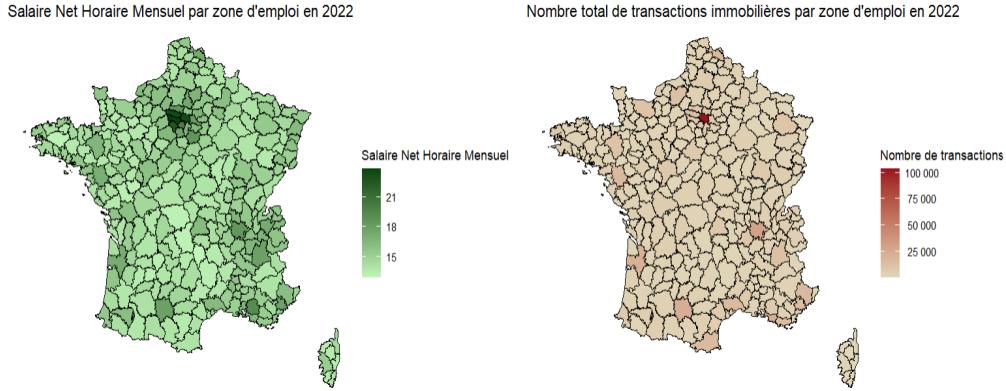


Figure 6: Carte ZE salaire et mutations

Désormais, à l'échelle des zones d'emploi, Paris se distingue très clairement par son nombre de transactions immobilières. Plus de 100 000 logements y ont changé de propriétaire en 2022, soit trois fois plus que le deuxième du classement : Lyon qui a suivi de Marseille, Toulouse et Bordeaux. La capitale confirme qu'elle est le coeur économique du pays avec une forte demande en logements. Malgré le fait que son offre immobilière soit limitée par le manque de place pour de nouvelles constructions, cela crée une forte rotation de par l'attractivité de ce territoire.

L'analyse descriptive des deux variables étudiées, à savoir les salaires horaires moyens et le nombre de transactions immobilières, met en lumière un possible lien entre ces indicateurs économiques. Les résultats montrent que les zones où les niveaux de salaire sont plus élevés tendent également à présenter un volume plus important de transactions immobilières, surtout dans la région parisienne. Cette relation pourrait traduire l'influence du pouvoir d'achat sur la capacité des ménages à investir dans le logement.

Par ailleurs, l'étude statistique des données révèle que des disparités salariales existent selon certains critères sociaux, ou bien géographiques. Les zones économiquement favorisées, souvent autour des grandes villes bénéficient d'une dynamique économique plus importante. Cela se traduit par un marché immobilier plus actif, compte tenu de la fréquence élevée de transactions. À l'inverse, dans les territoires où les salaires sont plus bas, nous observons généralement une plus faible intensité des transactions immobilières, signe d'une attractivité économique moins importante.

Cependant, cette analyse reste descriptive et ne permet pas à elle seule d'établir des liens entre les variables ni de mesurer précisément l'impact des transactions immobilières sur les salaires ou inversement. De plus, d'autres facteurs économiques, démographiques ou réglementaires tels que les

taux d'intérêt, l'offre de logement, ou encore la structure démographique locale peuvent également influencer la dynamique des transactions.

C'est pourquoi il apparaît nécessaire de recourir à des modèles économétriques, afin d'isoler l'effet propre de chaque variable observée. L'estimation de ces modèles permettra de mieux comprendre la relation entre les salaires et le nombre de transactions immobilières et constitue une étape essentielle pour approfondir un possible lien dynamique ou géographique.

### 3 Modèle Spatial

Dans notre étude économique, nous pouvons supposer que la position géographique d'une zone d'emploi en relation avec ses voisines peut influencer sur le niveau des salaires observés. En d'autres termes les salaires d'une zone d'emploi peuvent-ils être influencés par ceux des zones adjacentes ? Les caractéristiques des zones voisines permettent-ils d'expliquer les disparités salariales ?

Afin de répondre à ces questions, notre étude va se porter sur des modèles d'économétrie spatiale qui vont permettre d'intégrer les interactions géographiques entre nos observations.

Notre raisonnement va se dérouler en plusieurs étapes. Tout d'abord, nous allons étudier la relation entre les salaires et les mutations. Par la suite, nous élargirons notre champ d'analyse en intégrant, en plus de ces deux variables, d'autres variables que nous jugeons pertinentes (Dabet and Floch (n.d.)).

Tout au long de cette analyse, nous allons utiliser le log des variables car cela permet de réduire la variance de la distribution.

#### 3.1 Notions fondamentales

##### 3.1.1 L'autocorrelation spatiale

Le concept fondamental en économétrie spatiale est l'autocorrélation spatiale qui est une mesure de dépendance entre les valeurs de notre variable salaire sur différents lieux. Elle permet ainsi de déterminer si les zones d'emploi proches tendent effectivement à présenter des salaires similaires. Nous parlerons alors d'autocorrélation spatiale positive. Dans notre étude, nous pouvons nous attendre à une autocorrélation spatiale positive ,c'est-à-dire une concentration de zones d'emploi à hauts salaires autour d'autres zones elles aussi à hauts salaires et inversement pour les zones à bas salaires. Bien sûr, cette hypothese doit être vérifiée à travers un test.

##### 3.1.2 Matrice de Pondération Spatiale

Dans le premier temps de notre analyse, il convient de formaliser les relations spatiales entre les zones d'emplois à l'aide d'une Matrice de Pondération Spatiale. Cette matrice, que nous allons noter  $W$ , est un outil nécessaire pour modéliser les relations spatiales. Elle indique dans quelle mesure

deux zones d'emplois sont considérées comme voisines ou au contraire comme indépendantes. Nous allons pouvoir différencier plusieurs méthodes afin de construire cette matrice :

Matrice de contiguïté : les zones qui partagent une frontière commune sont alors considérées comme voisines, prenant la valeur 1 et les autres prennent la valeur 0. Cette matrice binaire sera normalisée par ligne.

Matrice de distances : la construction de cette matrice repose sur la distance entre les zones, qui sera souvent calculée à l'aide des centroïdes. Les poids sont définis comme l'inverse de la distance, puis normalisés. Cette distance peut être calculée selon plusieurs méthodes parmi laquelle nous pouvons retenir la distance euclidienne entre les centroïdes pour deux zones d'emploi.

Matrice des K voisins les plus proches (KNN) : chaque zone est associée à ses k voisins les plus proches calculés avec une mesure de distance. Le choix de k est important et nous allons devoir l'itérer pour trouver le plus optimal. Si il est trop faible, il va négliger certaines interactions et si il est trop grand, son influence locale sera trop faible.

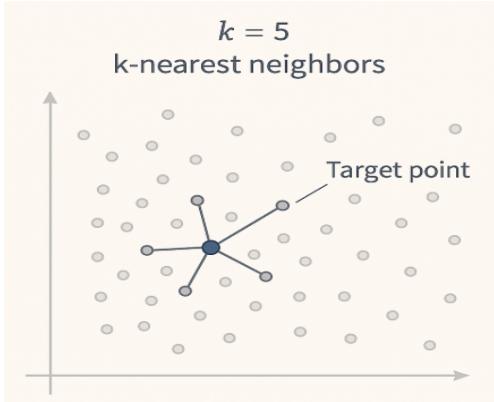


Figure 7: Explication méthode KNN

Nous avons choisi, dans un premier temps, de construire notre matrice de voisinage en utilisant la méthode de contiguïté. Il s'agit d'une matrice qui représente les relations spatiales entre unités géographiques : elle prend la valeur 1 si deux zones sont voisines, et 0 sinon.

Pour mieux visualiser ces relations spatiales, nous avons construit la carte ci-dessous, qui illustre les liens de proximité entre les zones d'emploi françaises. Ce graphique repose sur la matrice de contiguïté W. Chaque point représente une zone d'emploi, et les lignes rouges relient les zones considérées comme voisines. Sur le territoire représenté, chaque zone d'emploi compte au moins un voisin, au plus neuf, avec une moyenne d'environ cinq voisins par zone.

### **Voisinage entre zones**

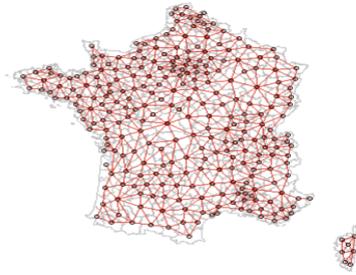


Figure 8: Voisinage entre Zone d'emplois

#### **3.1.3 L'indice de Moran**

Pour mesurer l'autocorrélation spatiale des salaires, nous allons pouvoir utiliser l'indice de Moran ,c'est un indicateur souvent employé pour tester la dépendance spatiale. (Moran (1950))

Cette mesure varie allant de -1 (autocorrélation négative) à 1 (autocorrélation positive) en prenant également la valeur 0 montrant l'absence d'autocorrélation spatiale. Cet indice compare la valeur obtenue par une zone à la moyenne pondérée des valeurs de ses voisines. Un test statistique permet de vérifier la signification de l'autocorrélation détectée, la distribution de la statistique de test peut être obtenue par des simulations de type Monte-Carlo.

$$\begin{cases} H_0 : \text{Absence d'autocorrélation spatiale(distribution aléatoire)} \\ H_1 : \text{Présence d'autocorrélation spatiale(valeurs similaires localisées)} \end{cases}$$

À présent que nous avons détaillé notre structure spatiale, il est nécessaire de tester une éventuelle présence d'autocorrélation spatiale à l'aide du test de Moran.

Ce test est effectué sur la variable salaire, qui constitue ici la variable dépendante. La p-value associée au test est égale à 2.2e-16, ce qui est extrêmement faible (voir Annexe 1). Pour un seuil de significativité  $\alpha = 5\%$  nous rejetons l'hypothèse nulle  $H_0$ , qui postule l'absence d'autocorrélation spatiale.

Il existe donc une forte autocorrélation spatiale globale positive des salaires (SNHM22) entre les différentes zones d'emploi.

Ces résultats justifient pleinement le recours à des modèles spatiaux, qui permettent de prendre en

compte l'interdépendance entre zones dans l'analyse des salaires. En effet, cette interdépendance viole les hypothèses classiques de l'OLS, rendant cette méthode inefficace et biaisée dans un tel contexte.

Maintenant que nous avons confirmé l'existence d'un effet spatial global significatif sur le salaire net horaire moyen (SNHM22), à l'aide du test de Moran global, nous cherchons à identifier dans quelles zones cet effet est localement significatif, et de quel type d'autocorrélation il s'agit.

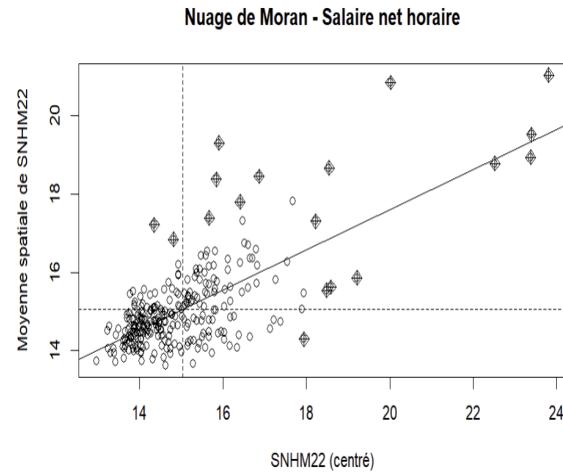


Figure 9: Nuage de Moran

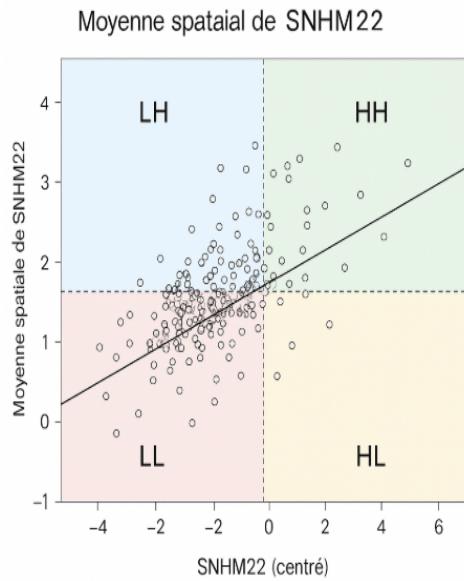


Figure 10: Exemple de representation d'un nuage de Moran

Le diagramme de Moran présenté s'est effectué sur nos données salariales avec une pondération

spatiale qui a été réalisée à l'aide d'une matrice de contiguïté. L'axe des abscisses indique la valeur centrée du salaire dans chaque zone d'emploi, tandis que l'axe des ordonnées affiche la moyenne spatiale de cette variable chez ses zones voisines, calculée à partir de la matrice de contiguïté.

Cette droite de régression montre une relation positive qui permet de constater que les zones d'emploi où les salaires sont élevés sont souvent entourées par des zones d'emploi avec des salaires similaires. Cela confirme le fait que les salaires élevés ou faibles ont tendance à se regrouper spatialement. Nous observons également quelques points dans les quadrants HL et LH, correspondant à des zones atypiques localement, qui peuvent être interprétées comme des anomalies spatiales (outliers positifs ou négatifs).

Cette représentation graphique renforce l'idée d'un effet spatial positif sur les salaires en France, confirmant la pertinence d'utiliser des modèles de régression spatiale dans l'analyse.

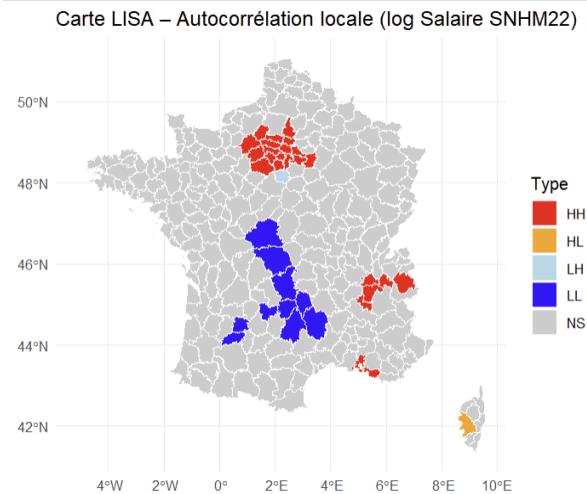


Figure 11: Autocorrelation Locale

Une carte peut être associée permettant de situer les zones d'emploi en fonction de leurs caractéristiques (HH signifie un salaire élevé dans un environnement élevé, HL un salaire élevé dans un environnement plus bas). Elle permet de constater que cette relation n'est pas homogène sur le territoire. Nous observons ainsi un cluster riche significatif en Ile de France et un cluster pauvre significatif en Auvergne.

## 3.2 Modèles SAR, SEM, SDM

### 3.2.1 Choix du modèle

Nous cherchons à expliquer les transactions en fonction des modèles estimés. L'hypothèse initiale que nous formulons à propos de cette relation est que les transactions n'expliquent pas nécessairement les salaires. En ce sens, nous nous attendons à obtenir des coefficients très faibles.

Par ailleurs, les résultats obtenus montrent que les erreurs du modèle OLS ne sont pas indépendantes spatialement, ce qui constitue une violation de l'hypothèse classique d'indépendance des erreurs dans les moindres carrés ordinaires (MCO). C'est pourquoi il est essentiel d'avoir recours à des modèles spatiaux tels que le SAR, le SEM ou encore le SDM.

Le modèle SAR: *Spatial Autoregressive Model* postule que la valeur de la variable dépendante  $y$  dans une unité géographique dépend directement des valeurs de  $y$  dans les unités voisines. Il modélise donc une interdépendance spatiale dans la variable expliquée.

Le modèle SEM: *Spatial Error Model* considère que l'autocorrélation spatiale provient d'omissions dans les variables explicatives, captées dans le terme d'erreur  $u$ . L'influence spatiale se manifeste donc dans les erreurs et non directement sur  $y$ .

Le modèle SDM: *Spatial Durbin Model* est une généralisation du modèle SAR. Il suppose à la fois une dépendance spatiale dans la variable dépendante  $y$  et dans les variables explicatives  $X$  des unités voisines (via  $WX$ ). Ce modèle permet ainsi d'analyser les effets directs qui sont propres à chaque zone d'emploi mais aussi les effets indirects provenant des zones voisines sur les niveaux des salaires.

$$Y = \rho WY + X\beta + WX\gamma + \varepsilon$$

où :

- $Y$  est le vecteur des salaires moyens par zone d'emploi,
- $WY$  représente l'influence des salaires voisins (effet spatial sur la variable dépendante),
- $X$  contient les variables explicatives locales (propres à chaque zone),
- $WX$  contient les mêmes variables, mais pour les zones voisines (effets indirects),

- $\rho, \beta, \gamma$  sont les coefficients à estimer,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$  est le terme d'erreur, supposé normal et homoscédastique.

Table 2: Coefficient,  $R^2$  et AIC pour chaque modèle

Statistique	OLS	SAR	SEM	SDM
Coefficient	0.0616	0.0476	0.0521	0.0528
R.	0.3030	0.6042	0.6428	0.6430
AIC	-636.8000	-763.6700	-778.7800	-778.1700

Figure 12: Comparaison des 3 modèles

Nous avons donc estimé plusieurs modèles spatiaux. Les coefficients associés à  $\log(\text{transactions})$  sont tous significatifs. Cette analyse permet de comparer les 3 modèles.

Comme anticipé, le modèle OLS présente la plus faible qualité d'ajustement, avec la valeur de l'AIC la plus élevée et le  $R^2$  le plus bas. Le modèle SAR améliore légèrement ces résultats, mais ce sont les modèles SEM et SDM qui affichent les meilleures performances, avec des AIC plus faibles et des  $R^2$  plus élevés.

Se pose alors la question du choix du modèle. Étant donné que le SDM est le modèle le plus général (il englobe SAR et SEM comme cas particuliers), nous avons souhaité vérifier s'il pouvait être simplifié en un modèle SAR. Pour cela, nous avons effectué un test du rapport de vraisemblance (LR test) avec les hypothèses suivantes :

$$\begin{cases} H_0 : \text{ le modèle restreint est vrai} \\ H_1 : \text{ le modèle non restreint est vrai} \end{cases}$$

Nous obtenons une p-value de 4.861e-05, qui est bien inférieure à 0.05. Cela signifie que, pour un seuil de 5 %, nous rejettons  $H_0$ . En d'autres termes, la variable explicative “log\_transaction” a un effet significatif sur la variable “log\_salaire”, et il est donc pertinent de conserver le modèle SDM. Ce modèle présente également une AIC plus faible et un  $R^2$  plus élevé que les autres modèles.

Le choix entre le modèle SDM et le modèle SEM est plus délicat. Nous avons d'abord effectué le test de Moran sur les résidus du modèle SDM : la p-value très élevée indique que nous ne rejettons

pas l'hypothèse d'absence d'autocorrélation, ce qui signifie que les résidus ne sont pas spatialement autocorrélés. Cela suggère que le SDM capture correctement la structure spatiale des données.

Cependant, nous avons également comparé les résidus du modèle SDM à ceux du modèle SEM (Annexe 3). Nous trouvons une corrélation très proche de 1 ( $\rho = 0.9975$ ), ce qui montre que les résidus des deux modèles sont très similaires. De plus, leurs AIC sont proches. Malgré cela, le SDM reste un choix cohérent, car il permet de prendre en compte à la fois les effets directs et indirects.

Nous avons pris la décision de garder le modèle SDM car il est particulièrement adapté à notre problématique, il permet ainsi de capturer aussi bien les effets des salaires des zones voisines (les effets spatiaux sur la variable dépendante) mais également les effets des caractéristiques des zones voisines (effets spatiaux sur les variables explicatives).

### 3.2.2 Application et Résultats

Nous avons estimé trois versions du modèle SDM à l'aide des matrices de pondération construites par différentes méthodes : contiguïté, KNN, distance. Le choix du nombre de voisins optimal s'est porté à 5 et à 8, (Annexe 4, Annexe 5) qui a été sélectionné avec le critère AIC.

Table 4: Coefficients estimés et diagnostics du modèle SDM

Variable	Matrice de voisinage			
	KNN 5	KNN 8	Contiguïté	Distance
(Intercept)	0.4238*** (1e-04)	0.3765** (0.003)	0.3897*** (0.00034)	0.3826** (0.0016)
log_trans	0.001(0.93)	0(1)	0.0086(0.49)	-0.001(0.93)
log_emploi	0.0477*** (1.2e-06)	0.0504*** (7.4e-07)	0.0434*** (3.9e-05)	0.0528*** (8.5e-08)
Chomage	-0.0114*** (7.7e-06)	-0.0126*** (4.4e-06)	-0.0139*** (3.2e-08)	-0.0122*** (7.9e-06)
W.log_trans	-0.0047(0.8)	0.007(0.75)	-0.0152(0.45)	-0.0076(0.68)
W.log_emploi	-0.0073(0.67)	-0.018(0.36)	-0.0087(0.64)	-0.0011(0.95)
W.Chomage	0.0085*(0.014)	0.0094*(0.017)	0.0125*** (0.00029)	0.0084*(0.027)
Rho	0.698	0.7168	0.7391	0.6845
\$R^2\$	0.7201	0.6942	0.7153	0.677
AIC	-846.04	-831.26	-832.8	-809.6

Table 1: Coefficients estimés du modèle SDM

Variable	Matrice de voisinage			
	KNN 5	KNN 8	Contiguïté	Distance
Constante	0.5487*** (7.1e-07)	0.4891*** (0.00018)	0.5427*** (1.2e-06)	0.5766*** (2.3e-06)
log(Trans)	0.0498*** (0)	0.0516*** (0)	0.0528*** (0)	0.0528*** (0)
W log(Trans)	-0.0196* (0.016)	-0.0243* (0.012)	-0.0314*** (1.8e-05)	-0.0201* (0.025)
Rho	0.7077	0.7379	0.737	0.6908
\$R^2\$	0.6559	0.6229	0.643	0.5954
AIC	-795.44	-779.12	-778.17	-754.32

Le modèle SDM basé sur la matrice des 5 voisins les plus proches a ici donné les meilleurs résultats (meilleur coefficient  $\rho$  pour l'ajustement global), ce qui en fait donc le modèle qui a été retenu pour la suite de notre analyse. Ce modèle permet donc de mieux comprendre les dynamiques salariales qui sont structurées géographiquement notamment avec les effets des zones d'emplois qui se partagent.

### 3.2.3 Interprétation Économique

À travers les différentes sorties réalisées en amont nous pouvons constater que la réalisation du modèle SDM expliquant le log(salaire) par le log(transaction) a donné un effet direct positif et significatif démontrant qu'une hausse des mutations dans une zone d'emploi sera corrélée à une hausse des salaires (attractivité économique). En revanche, son effet indirect (spatial) est significatif et négatif indiquerait qu'une augmentation des transactions dans les zones voisines concurrencerait la zone locale (les zones voisines prennent une part du dynamisme). À l'inverse lorsque nous intégrons plusieurs autres variables explicatives dans notre modèle SDM force est de constater que la variable log(transaction) perd en significativité démontrant l'effet qui a été purgé pour cette variable. Cette variable est donc la combinaison de plusieurs variables explicatives qui à elle seule perd en influence.

L'analyse économétrique vise à tirer parti du lien entre le salaire dans les zones d'emploi à travers les mutations. En parallèle de cela deux variables structurelles ont été introduites dans le modèle soit le chômage et la densité d'emploi afin de mieux capter les dynamismes socio-économiques locaux. Parmi les 4 configurations établies, nous pouvons remarquer que tous les coefficients de dépendance spatialen(rho) sont élevés et significatifs.

Cela confirme notre hypothèse initiale de présence d'autocorrélation spatiale dans les salaires selon laquelle des niveaux de salaires similaires tendent à se regrouper géographiquement. De plus, en tirant profit de ces informations nous pouvons constater que les inégalités salariales s'inscrivent dans des dynamiques régionales. Cette observation reste cohérente avec la géographie économique notamment avec l'idée selon laquelle le marché du travail s'étend au-delà des frontières administratives.

Nous pouvons constater que le log du nombre de mutations qui reflète la pérennité résidentielle et économique d'une zone d'emploi n'est jamais significatif dans les 4 modèles estimés avec un effet positif proche de zéro mais non significatif. Cette absence de significativité peut être déterminée par un effet marginal moins puissant dans le modèle ou encore que les mutations reflètent plutôt la mobilité résidentielle à l'affluence économique. De même, son effet spatial est également non significatif traduisant le fait que les dynamiques des mutations voisines n'influencent pas directement les salaires locaux.

La variable log emploi représentant le logarithme du nombre d'emplois dans la zone est hautement significative et positive pour l'ensemble des modèles. Ce résultat traduit un effet d'agglomération économique avec l'idée que plus une zone concentre des emplois alors plus la zone d'emploi est susceptible de proposer des salaires élevés. Ce phénomène se justifiant par une augmentation de la productivité et a pour but d'attirer des entreprises à plus forte valeur ajoutée. En revanche son effet spatial est non significatif montrant que les zones voisines n'ont pas d'influences significatives via leur niveau d'emploi confirmant ainsi que les effets d'agglomération sont plutôt locaux.

La variable chômage quant à elle est systématiquement négative mais également significative pour l'ensemble des modèles. Ce résultat est économiquement attendu en effet un chômage plus élevé aura tendance à tirer les salaires vers le bas et pouvant signaler une fragilité structurelle de l'économie locale. En revanche son effet spatial est positif et significatif ce qui signifie que le chômage dans les zones voisines augmente les salaires dans la zone locale. Un effet de rareté relative peut alors être envisagé où une zone peu touchée par le chômage devient attractive ce qui va augmenter la pression sur les salaires.

Le modèle basé sur la matrice KNN à 5 voisines présente le meilleur ajustement global car il possède l'AIC le plus faible et le  $R^2$  le plus élevé. Ainsi il capte de manière efficace les dynamiques spatiales proches et localisées. Cela montre l'importance de prendre en considération la dimension spatiale dans l'analyse des salaires qui révèlent que les zones à forte densité d'emploi rémunèrent mieux,

le chômage local freine les salaires et les mutations ne semblent pas un moteur des salaires. En somme, les caractéristiques des voisins influencent les rémunérations au-delà des effets locaux.

### 3.2.4 Variance Expliquée de notre modèle : Interprétation des effets spatiaux

La mesure des impacts dans le modèle SDM permet de quantifier de manière précise l'influence de nos variables explicatives en distinguant l'apport de l'effet propre à la zone d'emploi (effet direct) de celui provenant des zones d'emploi voisines (effet indirect). Ces deux effets combinés constituent l'effet total qui représente l'impact global d'une variable sur le salaire.

Le tableau suivant présente donc les effets estimés pour nos trois variables de notre modèle : le log du nombre de transaction, le log de l'emploi et le taux de chômage.

Table 1: Effets directs, indirects et totaux du modèle SDM

Variable	Effet direct	Effet indirect	Effet total
log(Transaction)	0.0002	-0.0122	-0.0120
log(Emploi)	0.0534	0.0803	0.1337
Chômage	-0.0113	0.0016	-0.0097

Figure 13: Effets directs, indirects et total

La variable  $\log(\text{transaction})$  permet de mesurer l'intensité des mutations dans une zone d'emploi ,son effet direct est quasiment nul suggérant que les transactions locales ont un faible impact sur les salaires locaux. Cependant son effet indirect est négatif indiquant que les mutations dans des zones voisines réduisent le salaire dans la zone locale. Cet effet peut-être associé à une concurrence territoriale : si des zones proches sont plus attractives alors cela peut diminuer la part de main d'œuvre qualifiée entraînant une pression à la baisse sur le marché du travail et ainsi sur les salaires. Son effet total est donc légèrement négatif ce qui renforce l'idée que les mutations surtout indirectes peuvent jouer dans la dynamique salariale.

Pour notre variable  $\log(\text{Emploi})$  elle mesure la densité d'emplois dans la zone qui présente des effets plus marqués dans le modèle. Son effet direct est positif montrant que pour une zone qui concentre beaucoup d'emplois alors cette zone d'emploi se verra proposer des salaires importants. Ce résultat suit la logique des effets où la concentration de l'activité économique est favorisée par

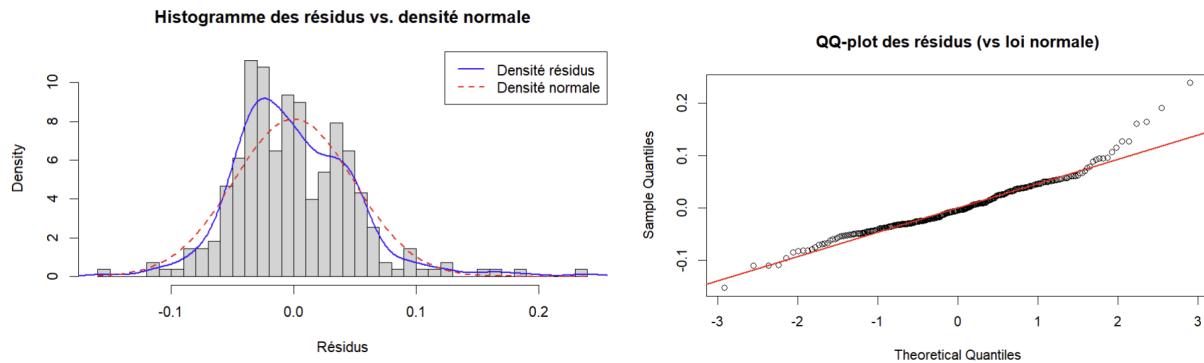
la productivité et les gains au niveau des salaires. De même il vient également un effet indirect plus important révélant que les zones d'emploi voisines avec une forte densité d'emplois augmentent les salaires locaux. Ce phénomène est tiré par des travailleurs qui peuvent vivre dans une zone d'emploi et travailler dans une autre zone d'emploi ou encore par des entreprises qui tirent profit des réseaux d'activités environnantes.

Enfin notre variable sur le taux de chômage agit comme un facteur mettant une pression à la baisse sur les salaires. Son effet direct est négatif signifiant que pour une zone d'emploi, si nous avons un chômage plus élevé il sera alors associé à des salaires plus faibles. Ce lien sous-jacent traduit un déséquilibre de l'offre et de la demande (plus de main d'œuvre limite les négociations salariales). À l'inverse son effet indirect est légèrement positif suggérant qu'un chômage de plus grande ampleur dans les zones voisines pourrait avoir une effet positif sur les salaires locaux. La zone d'emploi locale agirait comme un affluent économique bénéficiant aux entreprises et aux travailleurs en difficulté.

Ces résultats soulignent l'intérêt de la modélisation du modèle SDM qui permet ainsi de dissocier les effets propres d'une zone à ceux des zones voisines. Cela met en évidence l'intégration des interactions spatiales pour la compréhension des dynamiques salariales.

### 3.2.5 Analyse des Résidus du modèle SDM

Afin de vérifier la conformité de notre modèle spatial il convient d'analyser les résidus, dans le cadre de notre étude, la distribution des résidus du modèle SDM est réalisée à partir de la matrice de voisinage KNN( $k=5$ )



L'histogramme des résidus et de la densité de la loi normale montre que la distribution des résidus est plutôt centrée autour de zéro ce qui pourrait indiquer une absence de biais systématique dans les prédictions. En revanche la forme de la distribution a tendance à s'écartez de la loi normale

notamment avec plus de valeurs extrêmes. Notre analyse est confirmée à travers le QQ-plot dans lequel nous remarquons que la majorité des points suivent la droite mais les écarts s'accentuent aux extrémités montrant une déviation aux queues extrêmes par rapport à la distribution normale. Ces observations suggèrent que les résidus ne suivent pas une loi normale. Ce diagnostic se confirme avec un test de normalité de Shapiro-Wilk avec lequel nous observons une p-value inférieure à 1 (Annexe 6). Cela implique un rejet de l'hypothèse nulle de la normalité des résidus.

Nous constatons que les résidus ne sont pas normaux même si le fait qu'ils sont centrés autour de zéro indique que le modèle ne présente pas de biais dans ses prédictions salariales. Dit autrement le modèle ne surestime ni ne sous-estime les valeurs de la variable dépendante.

Le test de Breusch-Pagan sur les résidus du modèle SDM est réalisé avec la matrice des 5 voisins.

$$\begin{cases} H_0 : \text{Homoscédasticité des résidus (la variance des erreurs est constante)} \\ H_1 : \text{Hétérosécédasticité des résidus (la variance dépend des variables explicatives)} \end{cases}$$

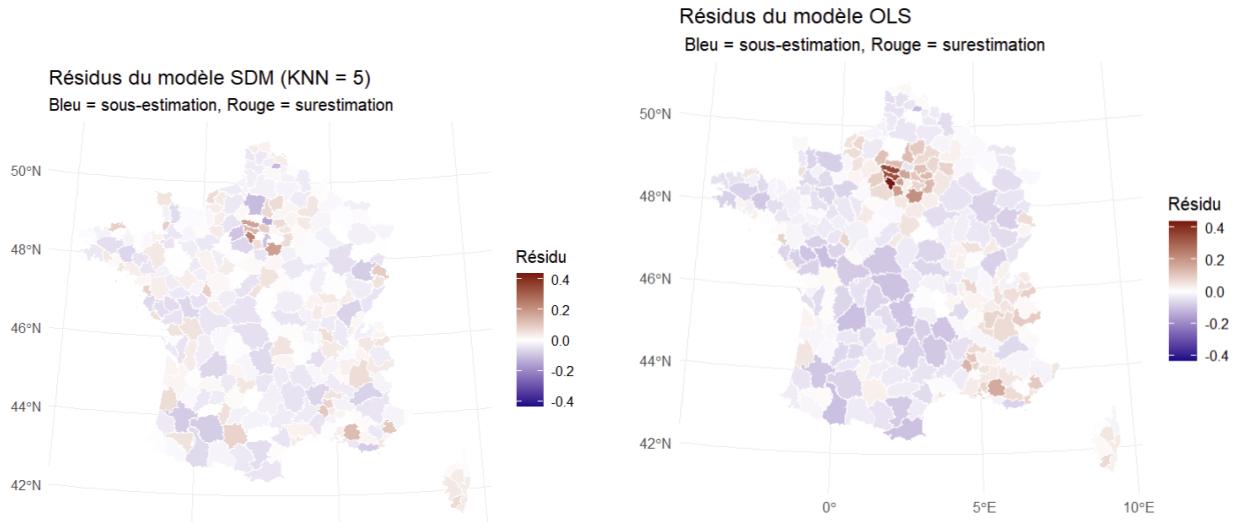
Table 5: Test de Breusch-Pagan sur les résidus du modèle SDM  
(KNN=5)

Modèle	Statistique	Degrés de liberté	p-value	Conclusion
SDM (KNN = 5)	6.0408	3	0.1096	Pas d'hétérosécédasticité significative

Figure 14: Test de Breush-Pagan

Ce test permet d'évaluer la variance de nos résidus en fonction des valeurs prises des variables explicatives. Si nous fixons un seuil de 5% avec notre statistique de test qui présente une p-value de 10.96% alors nous ne pourrons pas rejeter l'hypothèse nulle d'homoscédasticité. Ainsi les erreurs du modèle ne présentent pas de variance liée nécessairement aux variables explicatives. Cette absence d'hétérosécédasticité suggère que notre modèle ne possède pas de biais de dispersion ce qui renforce la robustesse des estimations et la fiabilité de nos conclusions économiques.

Cartographie des Résidus :



L’analyse spatiale des résidus nous permet de visualiser la qualité des prédictions de notre modèle au niveau national. Nous avons donc comparé les résidus issus du modèle linéaire avec ceux du modèle SDM toujours estimé à partir d’une matrice de voisinage KNN.

Sur la carte des résidus du modèle OLS, nous pouvons visualiser une concentration spatiale qui montre la présence d’autocorrélation spatiale qui n’a pas été prise en compte. Ce modèle néglige ici la dépendance entre zones voisines ce qui pourrait générer des erreurs localisées.

A l’inverse la carte des résidus du modèle SDM montre une nette amélioration avec des écarts qui sont moins extrêmes et moins concentrés géographiquement. Cela montre une réduction des zones avec des fortes erreurs et prouve que le modèle SDM corrige les biais spatiaux qui ont été constatés dans le modèle OLS.

La carte des résidus met en évidences que le modèle SDM surpassé le modèle OLS pour réduire les erreurs spatiales ce qui justifie pleinement le recours à l’économétrie spatiale. Le modèle spatial intègre des dépendances permettant de réduire l’autocorrélation spatiale dans les résidus.

### 3.3 Modèle GWR

#### 3.3.1 Modèle simple

Nous avons construit jusqu’ici des modèles spatiaux classiques comme les modèles SAR, SEM ou SDM. Ces modèles permettent de modéliser la dépendance spatiale structurelle à l’aide d’une matrice de voisinage. Cependant, ils reposent sur l’hypothèse d’effets globaux et homogènes dans l’espace, ce qui n’est pas toujours le cas en réalité.

C'est pourquoi nous utilisons ici un modèle GWR qui se différencie des modèles classiques par son approche locale de la relation entre les variables. Contrairement aux autres modèles, il permet de capturer l'hétérogénéité spatiale dans les relations entre la variable dépendante et les variables explicatives. Il ne suppose plus que l'effet est le même partout mais cherche au contraire à expliquer comment l'intensité de la relation varie selon le lieu.

Les résultats sont les suivants pour l'année 2022:

Coefficient	Min	X1er.Quartile	Médiane	X3e.Quartile	Max
Intercept	1.8261	2.1157	2.2330	2.3325	2.6884
log_transactions	0.0013	0.0468	0.0607	0.0737	0.1266

Figure 15: Résultat GWR

Comme précédemment, nous utilisons ici le modèle OLS à titre comparatif. Nous obtenons donc un coefficient associé au log(transaction) égal à 0,061 ce qui signifie qu'une hausse de 1 % du nombre de mutations immobilières est associée à une augmentation de 0,061 % du salaire net horaire.

Nous remarquons immédiatement que l'utilisation d'un modèle GWR augmente nettement le  $R^2$  : nous passons d'un  $R^2$  de 0,303 à 0,689 ce qui montre qu'il capte beaucoup mieux la structure spatiale du phénomène étudié comme c'était déjà le cas pour les modèles SAR, SEM et SDM. Les coefficients locaux du logarithme des mutations varient sensiblement selon les zones avec des valeurs comprises entre 0,0013 et 0,1266 pour une médiane de 0,0607. Cela signifie que selon la localisation, une même variation du volume de transactions n'a pas le même impact sur les niveaux de salaire.

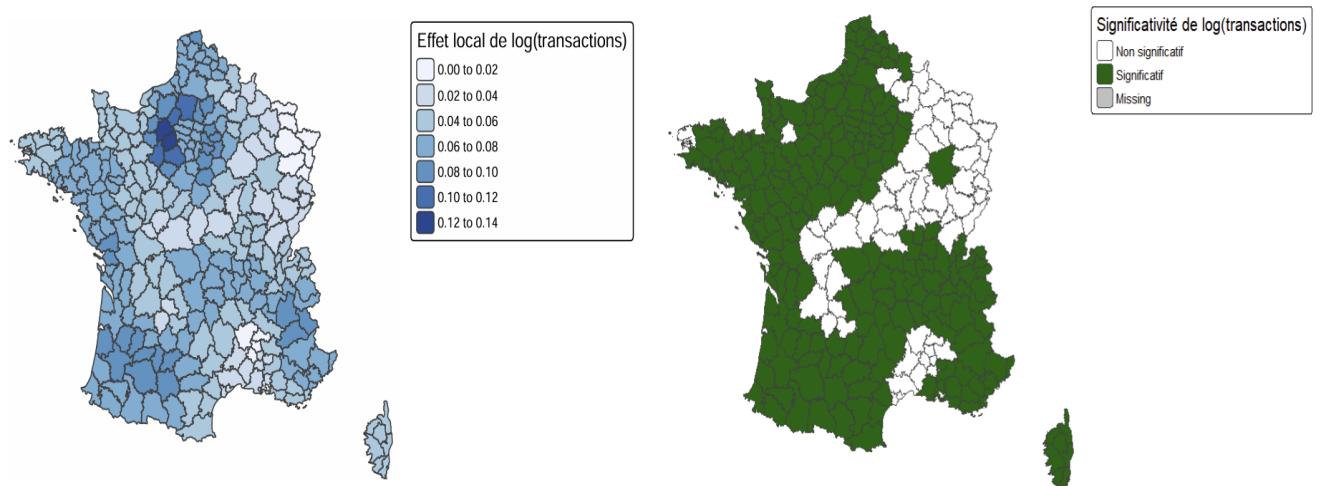


Figure 16: Carte des coefficient de log(transaction) et significativité

La carte représentant l'effet local de  $\log(\text{transactions})$  sur  $\log(\text{salaire})$  révèle une forte hétérogénéité spatiale de cette relation à travers les zones d'emploi françaises. D'un point de vue économique, les zones en bleu foncé, comme Paris, Toulouse, Nice, ou encore certaines zones du Sud-Ouest (notamment autour de Bordeaux) affichent les coefficients locaux les plus élevés. Cela suggère que dans ces territoires, une augmentation du nombre de mutations immobilières est associée à une hausse significative du salaire horaire net ce qui peut traduire des marchés dynamiques, une forte attractivité économique, ou encore une pression foncière accrue qui pourrait influencer les niveaux de rémunération. À l'inverse, dans de nombreuses zones plus centrales ou rurales (notamment en Bourgogne, Limousin ou Champagne-Ardenne), l'effet est beaucoup plus faible, indiquant que dans ces régions, l'activité immobilière joue un rôle secondaire dans la détermination des salaires. Ces disparités soulignent l'importance de prendre en compte la dimension spatiale dans l'analyse des mécanismes économiques locaux.

### 3.3.2 Modèle enrichi

Comme pour le modèle SDM, l'ajout d'autres variables comme le taux de chômage ou l'emploi fait que le coefficient associé au log des transactions perd en significativité pour une grande partie du territoire.

Tout de même, pour les quelques zones qui restent significatives, nous pouvons interpréter les coefficient associés à  $\log(\text{transaction})$ .

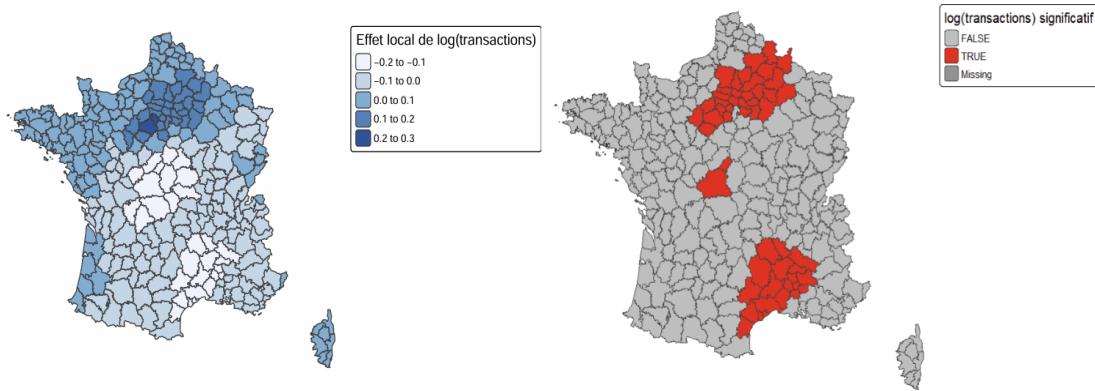


Figure 17: Effet local

Les zones d'emploi où la variable  $\log(\text{transactions})$  est statistiquement significative comprennent notamment l'Île-de-France, une partie de la Picardie et une partie du Centre-Val de Loire. Dans ces

régions, les coefficients estimés sont positifs compris entre 0,1 et 0,3 ce qui suggère qu'une hausse du nombre de mutations est associée à une augmentation significative des salaires. Ce résultat pourrait refléter une dynamique économique favorable : dans ces zones attractives et économiquement denses, un marché immobilier actif peut être le reflet d'un tissu économique dynamique, d'une demande forte en main-d'œuvre et donc de salaires plus élevés.

À l'inverse, certaines zones situées au sud-est du Massif Central présentent un effet significatif mais négatif du volume de mutations sur les salaires. Cela peut traduire une situation où une hausse des mutations immobilières n'est pas nécessairement le signe d'une dynamique économique vertueuse. Par exemple, cela pourrait être lié à une pression sur le foncier (spéculation, résidences secondaires) dans des zones moins productives ou à une migration résidentielle sans véritable création d'emplois qualifiés. Ces contrastes confirment l'importance de recourir à des modèles locaux comme le GWR qui permettent de capturer ces disparités spatiales fines dans les relations économiques.

Ces mêmes tendances spatiales étaient déjà visibles dans le modèle GWR simple incluant uniquement  $\log(\text{transactions})$  comme variable explicative. Nous observions alors des coefficients plus élevés en Île-de-France (allant de 0,10 à 0,14) et plus faibles dans les zones du sud-est du Massif central (autour de 0 à 0,04). Dans le modèle étendu, dans lequel sont intégrées également les variables  $\log(\text{emploi})$  et  $\text{chômage}$ , ces effets restent localement significatifs mais les coefficients associés à  $\log(\text{transactions})$  diminuent globalement en intensité.

Cette baisse des coefficients peut s'expliquer par le fait que les effets précédemment attribués uniquement aux mutations immobilières sont désormais partagés avec les nouvelles variables explicatives. L'introduction de ces variables dans le modèle permet donc de mieux isoler l'effet propre du marché immobilier, ce qui réduit mécaniquement l'amplitude des coefficients associés à  $\log(\text{transactions})$ . Cela renforce la pertinence du modèle enrichi qui offre une lecture plus nuancée et économiquement réaliste des déterminants spatiaux des salaires.

### 3.3.3 Analyse des résidus

Modèle simple :

Le graphique ci-dessous met en évidence une déviation notable de la normalité des résidus du modèle GWR. Bien que la forme générale de l'histogramme semble proche d'une distribution normale, une asymétrie légère ainsi qu'une queue étirée à droite suggèrent une distribution non parfaitement

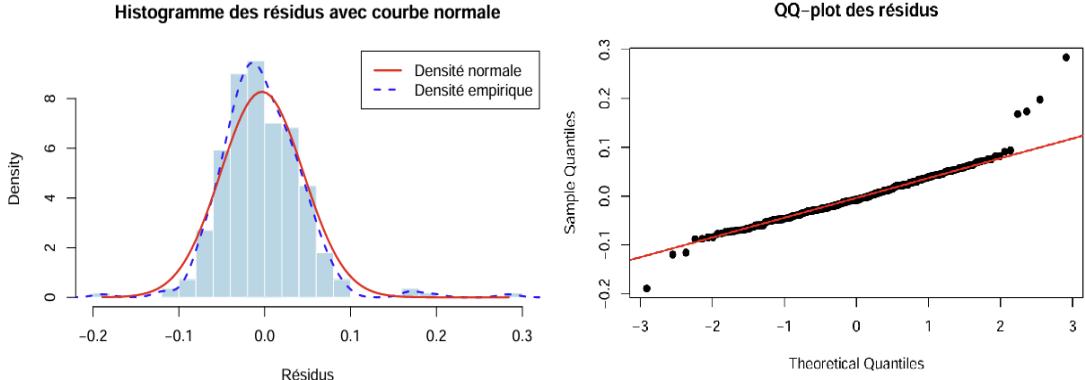


Figure 18: Résidus

normale. Ces observations sont confirmées à la fois par le test de Shapiro-Wilk et par le QQ-plot qui montrent des écarts significatifs aux extrémités.

Par ailleurs, une hétérosécédasticité modérée est détectée à travers la régression des carrés des résidus sur  $\log(\text{transactions})$ . Cela invite à une certaine prudence dans l'interprétation des coefficients locaux dont la significativité peut être affectée par l'instabilité de la variance des erreurs (résultats en annexe).

Modèle enrichi :

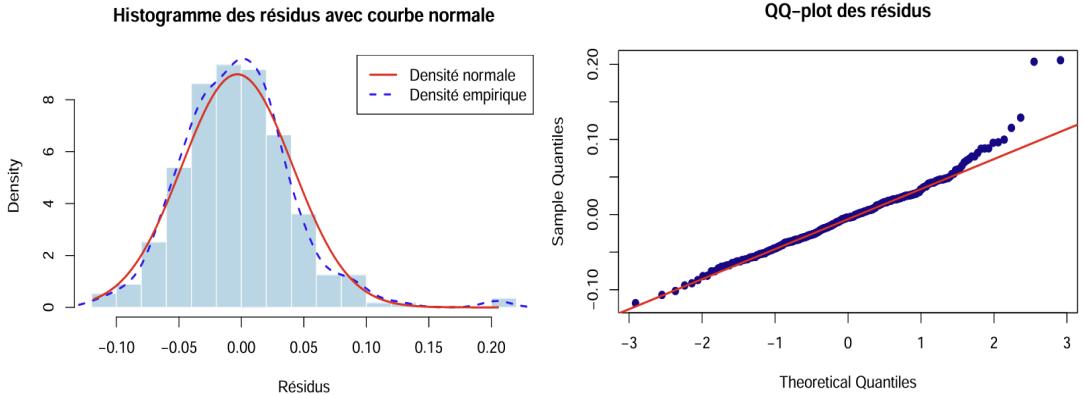


Figure 19: Résidus

Dans le modèle intégrant les variables  $\log(\text{emploi})$  et chômage, l'analyse graphique des résidus (histogramme et QQ-plot) ainsi que le test de Shapiro-Wilk confirment la non-normalité persistante des erreurs, notamment dans les valeurs extrêmes. Toutefois, la régression des résidus au carré sur les trois variables explicatives ne révèle aucune hétérosécédasticité significative, ce qui traduit une variance des erreurs plus homogène.

Cette amélioration s'explique par le fait que le modèle simple ne prenant en compte que log(transactions) omet des variables clés susceptibles d'expliquer à la fois les salaires et la dispersion des erreurs. En intégrant ces facteurs supplémentaires, le modèle devient plus complet et mieux spécifié ce qui permet de réduire la variabilité inexpliquée. Nous observons ainsi une diminution de l'hétérosécédasticité initialement présente renforçant la robustesse globale du modèle GWR enrichi (résultats en annexe).

### 3.4 Analyse complémentaire

#### 3.4.1 Transactions expliquées par les salaires

N'ayant pas obtenu de résultats concluants lors de nos premières estimations, nous avons pris la décision d'inverser la relation analysée. Nous avons donc cherché à examiner dans quelle mesure les transactions immobilières pouvaient être expliquées par les salaires moyens, ainsi que par d'autres variables explicatives d'ordre socio-économique.

Les tableaux ci-dessous présentent les résultats de ces nouvelles estimations à l'aide d'un modèle spatial Durbin (SDM), en tenant compte de la structure socio-professionnelle des zones d'emploi, du salaire moyen, et du taux de chômage.

Table 5: Coefficients estimés et diagnostics du modèle SDM

Variable	Matrice de voisinage			
	KNN 5	KNN 8	Contiguïté	Distance
(Intercept)	-0.8646(0.54)	-0.3533(0.82)	-0.0222(0.99)	-0.1978(0.9)
log_salaire	6.5913***(0)	6.3773***(0)	6.6672***(0)	6.3161***(0)
W.log_salaire	-4.6098***( $9.1e-09$ )	-4.7752***( $4.6e-09$ )	-4.9993***( $2.2e-09$ )	-4.3852***( $3.4e-07$ )
Rho	0.4356	0.5011	0.4325	0.3679
\$R^2\$	0.432	0.4191	0.4277	0.3938
AIC	538.84	543.79	541.89	553.27

Notre analyse permet de constater un coefficient direct de log(salaire) fortement positif et significatif indiquant que dans une zone locale, une hausse des salaires sera associée à une forte augmentation de changement de propriétaires. Un haut salaire peut refléter un environnement économique sain ce qui pourrait stimuler les changements résidentiels. Son effet indirect de log(salaire) est également significatif et négatif montrant que si des zones voisines tendent à avoir des salaires élevés cela

diminuerait les transactions locales. Si les zones voisines ont des salaires élevés, elles pourraient attirer les ménages au détriment de la zone locale. De plus les ménages peuvent préférer déménager vers des zones voisines qui rémunèrent mieux et cela pourrait réduire la pression sur le marché immobilier et dégrader l'attractivité immobilière locale. Au premier abord, nous pourrons montrer que le salaire est un facteur des transactions immobilières locales avec des conditions économiques des territoires voisins jouant un rôle de concurrence importante.

Table 3: Coefficients estimés et diagnostics du modèle SDM

Variable	Matrice de voisinage			
	KNN 5	KNN 8	Contiguïté	Distance
(Intercept)	-0.6233(0.31)	-0.5174(0.42)	-0.3477(0.57)	-1.0781(0.14)
log_salaire	-0.0084(0.98)	0.0636(0.84)	0.1715(0.58)	-0.0739(0.82)
log_emploi	0.8016***(0)	0.8016***(0)	0.7942***(0)	0.8055***(0)
Chomage	0.0609***(8.3e-06)	0.0689***(9.6e-07)	0.064***(4.6e-07)	0.0656***(5.7e-06)
lag.log_salaire	0.091(0.82)	-0.2921(0.49)	-0.1537(0.72)	0.1519(0.74)
W.log_emploi	-0.4994***(1.6e-15)	-0.4788***(1.2e-10)	-0.5473***(0)	-0.3803***(7.2e-07)
W.Chomage	-0.0442*(0.017)	-0.0637**(0.0016)	-0.0533**(0.002)	-0.0544**(0.0065)
Rho	0.6206	0.6946	0.6886	0.5148
\$R^2\$	0.8999	0.8982	0.9106	0.8869
AIC	77.74	79.2	56.32	102.11

Nous pouvons constater que le coefficient direct associé à log(salaire) n'est jamais significatif pour les 4 matrices de voisinages. Dans ce modèle le salaire horaire n'a pas d'effet direct sur le niveau des mutations immobilières dans une zone d'emploi. Une première intuition suggère que ce ne sont pas les salaires qui visent à expliquer les transactions immobilières.

Les autres résultats permettent de constater que la densité d'emploi est un déterminant important des transactions immobilières dû à sa significativité et son fort coefficient. Cela confirmerait le rôle des bassins économiques dans les dynamiques des résidences. De plus son effet indirect négatif peut être vu comme une concurrence territoriale où une zone voisine dynamique peut dissuader la mobilité de la zone locale.

La variable log(salaire) ne semble pas influencer directement ou même spatialement le niveau des transactions immobilières pouvant être traduit par le fait que les salaires influencent le type de logement mais pas le volume des transactions.

Quant à la variable chômage avec un effet direct positif et un effet indirect négatif, cela renforce notre idée d'une plus forte instabilité résidentielle ce qui peut provoquer des changement de résidences liés à la pression du marché immobilier. De plus, une zone d'emploi entourée d'un fort taux de chômage pourrait affaiblir la demande immobilière locale et donc générer un effet de déprime. Cela pourrait même envoyer un message fort aux investisseurs et aux ménages qui peuvent craindre une dévalorisation immobilière dans la zone locale. Les dynamiques immobilières sont dans un tissu géographique connecté ensemble.

Nos résultats montrent que le salaire, pris isolément, présente un lien significatif et positif avec le nombre de transactions immobilières. Toutefois, lorsqu'e nous introduisons dans le modèle d'autres variables structurelles comme le chômage ou la densité d'emploi, cet effet disparaît totalement. Ce résultat suggère que le salaire n'est pas, à lui seul, un déterminant direct des mutations immobilières.

### 3.4.2 Etudes sur années antérieures

Variable	KNN 5	KNN 8	Contiguïté	Distance
Constante	0.4938*** (2.1e-07)	0.4533*** (6e-05)	0.4261*** (3.8e-06)	0.4785*** (3.9e-06)
log(Trans)	0.0332*** (0)	0.0357*** (0)	0.0378*** (0)	0.0364*** (0)
W log(Trans)	-0.0102 (0.086)	-0.0115 (0.12)	-0.0254*** (8.8e-06)	-0.0086 (0.23)
Rho	0.7396	0.7521	0.7975	0.733
R <sup>2</sup>	0.7128	0.6769	0.6945	0.6397
AIC	-869.61	-846.81	-838.37	-806.69

Figure 20: Année 2014

Matrice de voisinage		
Variable	KNN 5	Distance
Constante	0.5133*** (1.8e-07)	0.468*** (1.2e-05)
mutations	0.0337*** (0)	0.0369*** (0)
lag.mutations	-0.011 (0.071)	-0.0115 (0.12)
Rho	0.7327	0.7432
R <sup>2</sup>	0.6934	0.6201
AIC	-841.57	-783.73

Figure 21: Année 2016

Après avoir représenté l'estimation des modèles SDM pour les années 2014 et 2016, cela met en évidence le lien significatif entre la variable log(mutations) et log(salaire). Ce résultat robuste aux

différentes matrices de voisinage confirme qu'une hausse de l'activité immobilière locale est associée à une augmentation des salaires horaires moyens. En revanche, les effets indirects c'est-à-dire ceux des zones voisines ne sont significatifs négativement que pour la matrice de contiguïté en 2014. Cela pourrait traduire un effet de réallocation de l'activité entre les territoires limitrophes.

Les résultats obtenus pour l'année 2022 confirment également cette relation directe positive entre les mutations immobilières et les salaires montrant la persistance de ce phénomène dans le temps. Le lien observé s'inscrit dans une logique durable où les zones d'emploi connaissant les mutations les plus importantes restent structurellement celles où les salaires sont les plus élevés. La stabilité de ce lien au fil du temps s'explique par l'inertie des dynamiques immobilières propres aux zones d'emploi ainsi que par le rôle des mobilités résidentielles dans l'évolution du marché du travail. Malgré le changement des zones d'emploi en 2020, nous retrouvons des résultats similaires dans les années antérieures.

## 4 Panel sur Zones d'emploi

De par le changement conséquent des zones d'emploi entre les années 2010 et 2020, nous nous sommes dirigés dans un premier temps sur une approche panel sur les départements, cependant au vu des résultats peu satisfaisants (Annexe Panel), nous avons finalement opté pour du panel spatial sur les zones d'emploi. Ce choix est justifié et motivé aussi par les résultats trouvés dans la section précédente concernant la stabilité des résultats sur le modèle spatial au cours du temps.

Nous considérons une analyse en panel spatial sur les zones d'emploi afin de prendre en compte l'approche dynamique qui manquait dans la précédente section. Notre analyse se base sur les 278 zones d'emploi 2020 dont nous disposons du nombre de logements ayant changé de propriétaire (base DVF). L'Alsace est manquante, comme indiqué précédemment. Les données sont observées de 2014 à 2022. Malgré la redistribution des zones d'emplois au milieu de notre période d'observation, le zonage 2020 sera celui sur lequel nous allons baser cette analyse.

Notre panel est ainsi composé de 2502 observations : 278 individus (zones d'emplois) sur 9 périodes (années) et nos variables étudiées sont le salaire net horaire moyen et le nombre de mutations.

### 4.1 Analyse préliminaire des données

Avant d'envisager une modélisation économétrique en panel avec dimension spatiale, nous commençons par une analyse descriptive de nos deux variables d'intérêt. Cette étape peut nous aider à mieux appréhender la distribution des salaires nets horaires moyens et du nombre de mutations immobilières au sein des différentes zones d'emploi et sur l'ensemble de la période considérée.

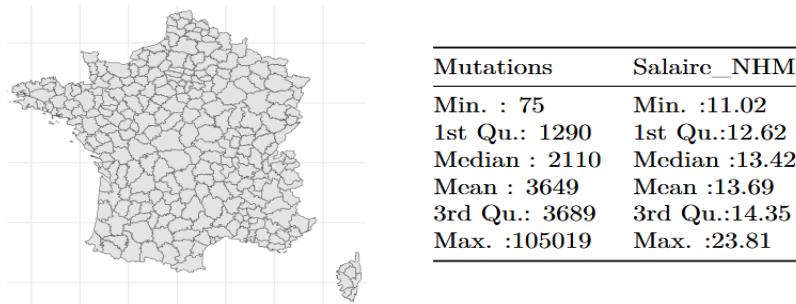


Figure 22: Stat. Des.

Concernant les mutations, leur nombre varie fortement entre les observations. La valeur minimale est de 75, tandis que la valeur maximale atteint plus de 105 000, ce qui suggère une forte

hétérogénéité entre les zones d'emploi. La moyenne s'établit à 3 649 mutations, mais la médiane est nettement plus basse (2 110), signe d'une distribution asymétrique avec quelques zones très dynamiques tirant la moyenne vers le haut comme nous l'avions constaté lors de la représentation géographique de nos données avec notamment Paris qui comptait 3 fois plus de mutations que le deuxième du classement. Le premier et le troisième quartile (1 290 et 3 689 respectivement) confirment que la majorité des zones présentent un volume de transactions modéré.

En ce qui concerne le salaire net horaire moyen, les écarts sont moins marqués, bien qu'une certaine dispersion soit observable. Le salaire horaire moyen varie de 11,02 euros à 23,81 euros, avec une moyenne de 13,69 euros et une médiane de 13,42 euros. Les quartiles montrent que 50 % des observations se situent dans un intervalle relativement restreint, compris entre 12,62 euros et 14,35 euros. Cela reflète une structure salariale globalement homogène, mais avec la présence de quelques zones à haut revenu certainement en région parisienne comme nous l'avions constaté.

#### 4.1.1 Paramétrage de la structure spatiale

Afin de modéliser notre panel spatial, nous devons définir une matrice de poids, qui représente la pondération des relations spatiales dans le modèle, comme dans la section précédente. La Corse étant présente dans nos observations, nous utiliserons la matrice KNN comme matrice de poids et non la matrice de contiguïté. Nous avons pu établir le nombre optimal de voisins (12), en maximisant la log-vraisemblance. Ce nombre de voisins à prendre en compte paraît raisonnable compte tenu de certaines faibles superficies des zones d'emploi. Il est probable que les 12 zones d'emploi les plus proches de la zone d'intérêt influencent celle-ci.

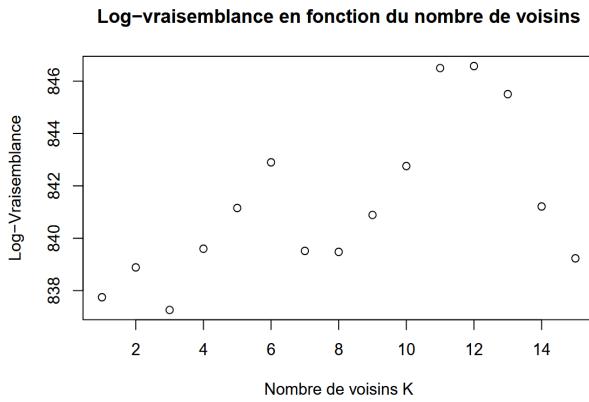


Figure 23: Choix Voisins K

Désormais, afin de nous assurer que notre approche spatiale soit justifiée, nous devons nous assurer qu'il existe de l'autocorrélation spatiale au sein de notre panel, et ce, pour chaque année. Pour ce faire, nous avons effectué un test de présence d'autocorrélation spatiale, le test de Moran. De plus, nous avons calculé la corrélation entre nos deux variables d'intérêt.

	P_Value	correlation
2014	0	0.5230959
2015	0	0.5243049
2016	0	0.5308843
2017	0	0.5185938
2018	0	0.5240487
2019	0	0.5313399
2020	0	0.5524765
2021	0	0.5512130
2022	0	0.5382724

Figure 24: Autocorrelation Spatiale et correlation

Le test de Moran nous confirme bien de la présence d'autocorrélation spatiale. Ainsi, les valeurs observées dans une zone d'emploi sont corrélées avec celles des zones voisines chaque année. Il est ainsi justifié de modéliser nos données à l'aide d'un panel spatial. De plus, les salaires nets horaires moyens et le nombre de logements changeant de propriétaire sont corrélés positivement. Cette corrélation semble stable de 2014 à 2022. Cela suggère qu'en moyenne, plus le salaire est élevé dans une zone, plus nous observons de transactions immobilières.

#### 4.1.2 Spécification du modèle

En vu de configurer notre modèle, nous devons déterminer la présence d'effets individuels et dans un second temps si ceux-ci sont corrélés aux explicatives. Nous avons mis en place des tests standards et robustes à l'autocorrélation spatiale.

Stat.F	P.Value	Test	Stat.H	P.Value
57.248	0	Standard	495.04	0
112.560	0	Robuste	59.08	0
		Lag	21.27	0

Figure 25: Test Fisher Hausman

Le test de Fisher nous confirme la présence d'effets individuels. Ces effets inobservables propres à chaque zone d'emploi les différencient structurellement des autres. Nous nous attendions à ce

résultat car chaque zone d'emploi a une différente qualité d'infrastructure ainsi qu'une attractivité naturelle ou bien un dynamisme du marché du travail différent. Ce sont autant de critères individuels et importants qui permettent de distinguer les zones d'emploi et qui sont importants à prendre en compte lors de la spécification de notre panel.

Le test de Hausman nous indique que les effets individuels sont fixes. C'est-à-dire que les caractéristiques spécifiques aux zones d'emploi sont corrélées au nombre de transactions. En adoptant un modèle à effets fixes, nous neutralisons statistiquement ces spécificités structurelles constantes dans le temps. Cela permet de mieux identifier l'effet propre du nombre de transactions immobilières sur les salaires indépendamment des particularités locales permanentes. Cependant, la transformation *within* va tendre à éliminer une grande partie de la variation spatiale persistante, notamment parce que les zones d'emploi connaissent relativement peu de changements structurels d'une année sur l'autre. Cela peut rendre l'interprétation de certains effets spatiaux plus délicate. Le recours à l'estimateur en erreurs aléatoires permettrait de préserver cette variation, mais supposer que les effets spécifiques aux zones d'emploi sont aléatoires et non corrélés aux variables explicatives semble économiquement discutable. En effet, des caractéristiques structurelles comme l'attractivité territoriale, les infrastructures ou la composition sectorielle sont difficilement assimilables à des effets aléatoires. Dans cette optique, le choix des effets fixes apparaît plus approprié pour capter la spécificité des territoires tout en garantissant la robustesse des résultats.

Ce choix méthodologique renforce la robustesse des résultats en limitant les biais d'omission dus à des facteurs inobservables fixes et garantit une interprétation économiquement plus fiable des liens dynamiques entre le nombre de mutations et les salaires au sein des zones d'emploi.

Par ailleurs, afin de décider du choix du modèle, nous avons réalisé différents tests LM robustes sur la dépendance spatiale de nos salaires (lag) et sur l'autocorrélation spatiale de nos erreurs.

Test	Stat.LM	P.Value
LM lag	1305	0
LM erreurs	1113	0

Figure 26: Test LM

Ces tests révèlent la présence à la fois d'un effet spatial lag, c'est-à-dire que la valeur des salaires dans une zone d'emploi dépend celle des zones voisines, mais aussi d'une autocorrélation spatiale dans les erreurs. Face à cette double présence de dépendance spatiale, ni un modèle SAR, ni

un modèle SEM n'est suffisant. Nous choisissons donc d'estimer un modèle SDM, comme dans la section précédente, qui permet de prendre en compte à la fois l'effet spatial de la variable dépendante et des variables explicatives.

## 4.2 Estimation

Nous avons ainsi spécifié un modèle panel qui vise à estimer l'impact du nombre de mutations immobilières (en milliers) sur le salaire net horaire moyen, tout en tenant compte de la dimension spatiale. Trois variantes de ce modèle ont été estimées : un modèle SDM pool (sans effets spécifiques ni temporels), un SDM avec effets fixes individuels mais sans effets temporels, et un SDM enrichi incluant à la fois des effets fixes spatiaux et temporels, afin de mettre en évidence l'importance de contrôler les effets fixes dans notre analyse spatiale des salaires.

Variables	SDM Pool	SDM sans effets temp	SDM
(Intercept)	13.261***	NA	NA
m_mutations	0.118***	0.06***	0.054***
lambda	NA	0.937***	-0.234***

Figure 27: Estimations

Ces résultats montrent le lien dynamique entre les mutations et les salaires. Le coefficient associé aux mutations est significatif et positif dans les trois estimations, confirmant la robustesse de l'effet des mutations immobilières sur le salaire net horaire moyen. En revanche nous observons une diminution progressive de ce coefficient à mesure que la spécification du modèle se complexifie. Dans le modèle SDM pool, le coefficient est estimé à 0.118 tandis qu'il chute à 0.06 dans le modèle avec effets fixes individuels puis à 0.054 dans le modèle SDM complet avec effets fixes temporels. Cette évolution est économiquement cohérente : le modèle pool, qui ne contrôle pas les effets spécifiques aux zones d'emploi ni les effets temporels, tend à surestimer l'effet des mutations immobilières sur les salaires. En effet, il attribue aux variations des mutations des effets qui proviennent en réalité de caractéristiques structurelles constantes dans le temps. Cela pourrait suggérer que les zones dynamiques tendent effectivement à offrir des salaires légèrement supérieurs, mais que cet effet est atténué par des facteurs locaux et temporels. Le modèle SDM complet, en neutralisant ces effets fixes, permet ainsi d'estimer de manière plus rigoureuse l'effet propre des mutations immobilières sur les salaires, à l'intérieur de chaque zone, au fil du temps. Ce dernier est plus fiable.

Une structure d'erreur spatialement autocorrélée est introduite dans les 2 derniers modèles, ce qui permet l'estimation de lambda, paramètre qui représente l'autocorrélation spatiale des erreurs. Il est alors estimé à 0.937 dans le modèle sans effets temporels, suggérant une forte autocorrélation spatiale des erreurs. Cependant, dans le modèle SDM enrichi, le lambda devient négatif et plus petit. Ce changement de signe et d'ampleur traduit le fait qu'une fois les effets fixes correctement pris en compte, la dépendance spatiale résiduelle dans les erreurs devient faible, voire inverse. Cela montre l'importance de contrôler les spécificités structurelles propres aux zones d'emploi pour ne pas faussement attribuer aux effets spatiaux des dynamiques qui relèvent en réalité d'une hétérogénéité non observée.

Ainsi, au vu des estimations de notre modèle SDM et toutes choses égales par ailleurs, une augmentation de 1 000 mutations immobilières dans une zone d'emploi donnée est associée à une hausse moyenne de 0,054€ du salaire net horaire moyen dans cette même zone. Ce résultat indique une relation positive entre le dynamisme du marché immobilier, caractérisé par le nombre de transactions et le niveau des salaires locaux. Cette relation peut s'expliquer par plusieurs facteurs économiques. Une hausse du nombre de transactions immobilières peut refléter un essor économique local, une attractivité croissante du territoire ou une demande accrue de main-d'œuvre, autant de phénomènes susceptibles d'exercer une pression à la hausse sur les salaires. Ce résultat suggère également que les marchés immobiliers et du travail sont interconnectés à l'échelle locale.

Le modèle SDM complet permet également de distinguer les effets directs, indirects, et totaux du nombre de mutations immobilières sur les salaires.

Table 2: Mesures d'impact

Variable	Direct	Indirect	Total
m_mutations	0.0537581	-0.0103354	0.0434227

L'effet direct est estimé significativement à 0.0538. Cela signifie que, toutes choses égales par ailleurs, une augmentation du nombre de mutations dans une zone d'emploi donnée entraîne une hausse du salaire net horaire moyen dans cette même zone. Ce résultat est économiquement compréhensible car une hausse des transactions immobilières peut refléter un dynamisme local, une augmentation de la demande de logement liée à l'emploi, ou une amélioration des conditions de vie, autant de facteurs susceptibles de soutenir ou renforcer le niveau des salaires.

L'effet indirect est quant à lui estimé significativement à -0.0103. Cela suggère que l'augmentation du nombre de mutations dans les zones voisines est associée à une baisse du salaire dans la zone considérée. Cela peut s'expliquer de plusieurs façons : il est possible que des effets de concurrence entre bassins d'emploi ou de mobilité résidentielle détournent des ressources ou des populations actives vers des zones d'emploi voisines, ce qui peut affaiblir localement le marché du travail.

Enfin, l'effet total qui est la somme des effets direct et indirect reste positif et fortement significatif, estimé à 0.0434, ce qui confirme que l'impact global du marché immobilier, même en tenant compte des interdépendances spatiales, est favorable au niveau des salaires.

Ces résultats confirment la pertinence du choix du modèle SDM avec effets fixes, justifié par les tests de spécification et économiquement interprétable. Il permet d'isoler l'effet des mutations immobilières sur les salaires en neutralisant les biais liés aux caractéristiques fixes des zones d'emploi. En outre, il capture la complexité des interactions spatiales, en distinguant les effets internes à chaque zone des effets des zones voisines. Ce niveau de précision analytique est essentiel dans une logique de diagnostic territorial et de formulation de politiques économiques adaptées aux dynamiques locales.

## 5 Dimension temporelle

Dans un contexte où les données évoluent continuellement dans le temps, l'analyse en séries temporelles s'impose comme un outil essentiel pour mieux comprendre et modéliser les dynamiques entre les séries. Cette approche permet de révéler les dépendances temporelles, les tendances ou les changements soudains susceptibles d'affecter significativement l'interprétation des phénomènes étudiés.

Dans cette partie, nous allons nous intéresser à la dimension temporelle de nos données afin de déceler l'existence d'une éventuelle dépendance dans le temps entre nos deux séries : le nombre de transactions et les salaires mensuels de base. Ces deux variables, dont nous pouvons soupçonner un lien économique potentiel, seront analysées sur la période du premier trimestre 2014 (T1 2014) au quatrième trimestre 2023 (T4 2023), soit un total de 40 observations trimestrielles pour cette modélisation temporelle.

Il convient de noter que les salaires mensuels de base ne sont disponibles qu'à une fréquence mensuelle et qu'ils sont exprimés en base 100, ce qui ne permet donc pas de connaître les niveaux réels de rémunération.

Enfin, toutes les séries seront exprimées en base 100 T1 2014 afin de garantir la cohérence et de faciliter les comparaisons et l'interprétation des évolutions au cours du temps.

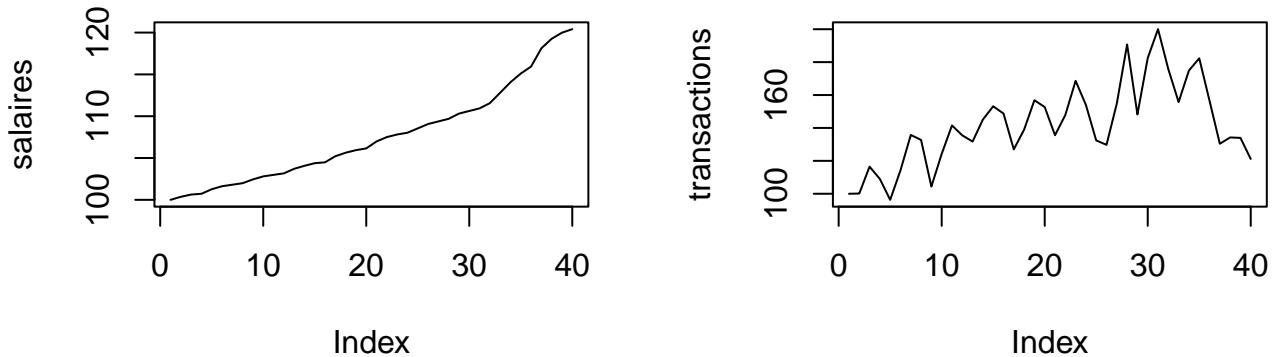
### 5.1 Analyse de la stationnarité des séries

Avant toute modélisation, il est nécessaire de vérifier la stationnarité de nos deux séries. Il est important qu'elles soient stationnaires car n'importe quelle régression linéaire sur des données temporelles non stationnaires donnerait lieu à une régression fallacieuse qui fausserait les résultats. Pour qu'une série soit stationnaire (faiblement), son espérance doit être constante au cours du temps, synonyme d'absence de tendance, sa variance doit également être constante et finie, et la covariance entre deux observations séparées d'un même écart temporel (lag) doit rester constante au fil du temps.

Vérifier la stationnarité constitue ainsi une première étape essentielle de notre démarche analytique. Elle garantit la validité des analyses temporelles qui suivront, en s'assurant que les propriétés statistiques des séries restent stables dans le temps.

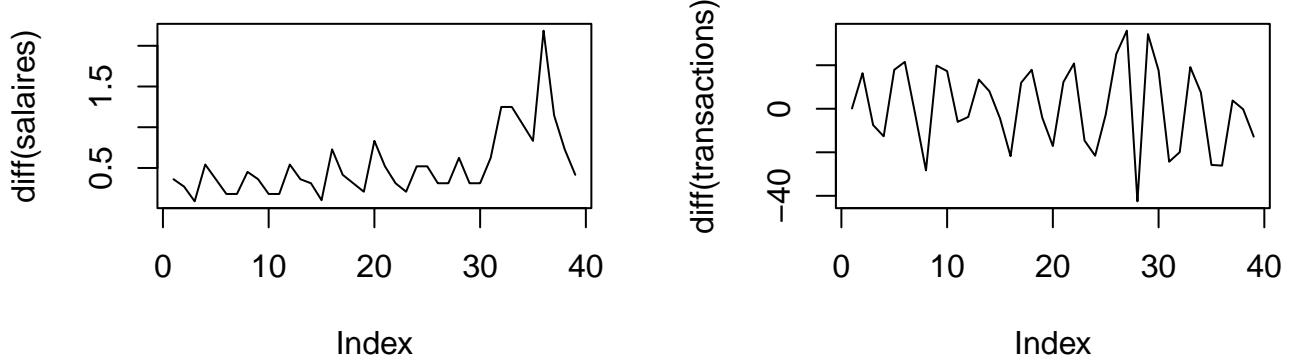
Nous pourrons ensuite étudier de manière fiable la relation entre le nombre de transactions immobilières et l'évolution des salaires, et ainsi déterminer s'il existe un lien temporel significatif entre ces deux variables, et si oui, quel en est le délai d'ajustement.

Nous commençons par une étude visuelle des séries en niveau :



L'observation des représentations graphiques des séries en niveau met en évidence une tendance haussière marquée dans les deux cas : à la fois pour le nombre de transactions, qui traduit une hausse globale de l'activité immobilière et pour les salaires mensuels de base, qui représente la croissance progressive des salaires. Cette dynamique visuelle suggère déjà une non-stationnarité, puisque l'hypothèse de moyenne constante au cours du temps ne semble pas respectée.

Afin d'observer l'effet d'une différenciation, nous représentons également les séries en différences premières :



Pour confirmer ces observations visuelles, nous avons réalisé une série de tests de stationnarité : KPSS (Kwiatkowski-Phillips-Schmidt-Shin), ADF (Augmented Dickey-Fuller) et PP (Phillips-Perron) dont voici les hypothèses :

KPSS :

$$\begin{cases} H_0 : \text{La série est stationnaire} \\ H_1 : \text{La série n'est pas stationnaire} \end{cases}$$

ADF et PP :

$$\begin{cases} H_0 : \text{La série n'est pas stationnaire} \\ H_1 : \text{La série est stationnaire} \end{cases}$$

Table 3: Résultats du test KPSS (type = ‘tau’)

Série	Statistique de test	Valeur critique à 5%	Stationnarité
Salaires (niveau)	0.232	0.146	Non
Transactions (niveau)	0.158	0.146	Non
Salaires (diff)	0.125	0.146	Oui
Transactions (diff)	0.113	0.146	Oui

Les résultats du test KPSS montrent qu'en niveau, nous pouvons rejeter l'hypothèse nulle de stationnarité au seuil de 5% : les statistiques de test dépassent les valeurs critiques. Les séries sont donc non stationnaires. En revanche, en différences premières, nous ne rejettons pas l'hypothèse nulle : les statistiques deviennent inférieures aux valeurs critiques, ce qui signifie que les séries sont devenues stationnaires après différenciation. On dit qu'elles sont intégrées d'ordre 1, noté I(1).

Les résultats des tests ADF et PP en annexe confirment globalement ces conclusions à une exception près. Le test PP confirme l'absence de stationnarité en niveau, puis la stationnarité après différenciation. Le test ADF, quant à lui, ne permet pas de conclure à la stationnarité des séries après différenciation. Cette différence peut s'expliquer par la sensibilité du test ADF et par la taille limitée de notre échantillon (40 observations).

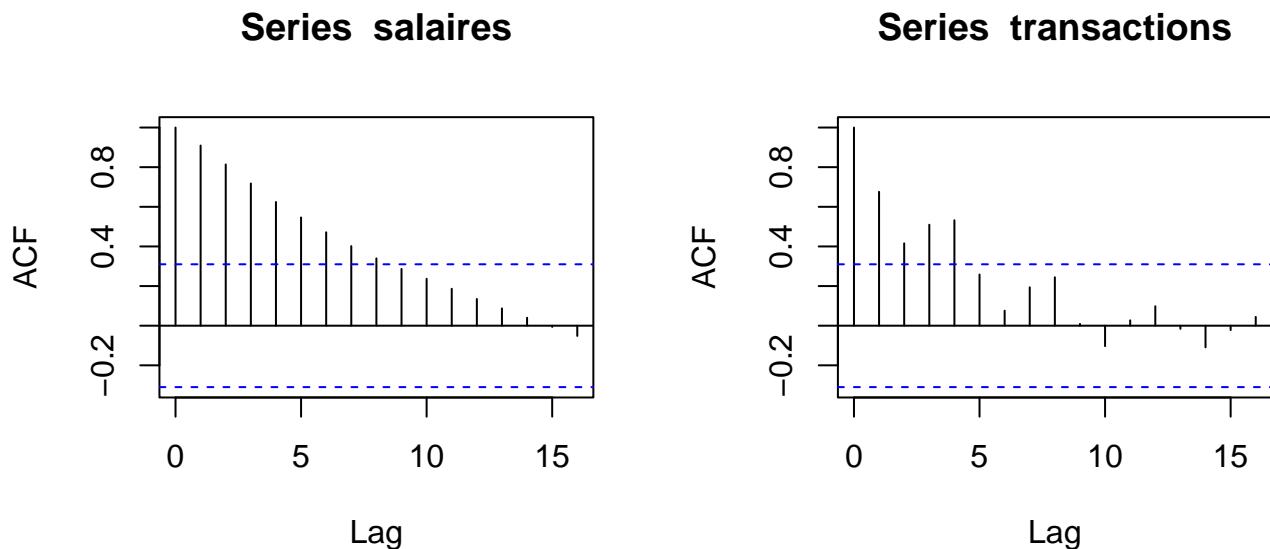
Au final, compte tenu des résultats convergents des tests KPSS et PP et du caractère incertain du test ADF, nous retenons que les séries sont intégrées d'ordre 1 : I(1). Cela est typique des variables économiques et déterminant pour la suite de notre analyse car nous pourrons tester l'existence d'une relation de cointégration entre les deux séries.

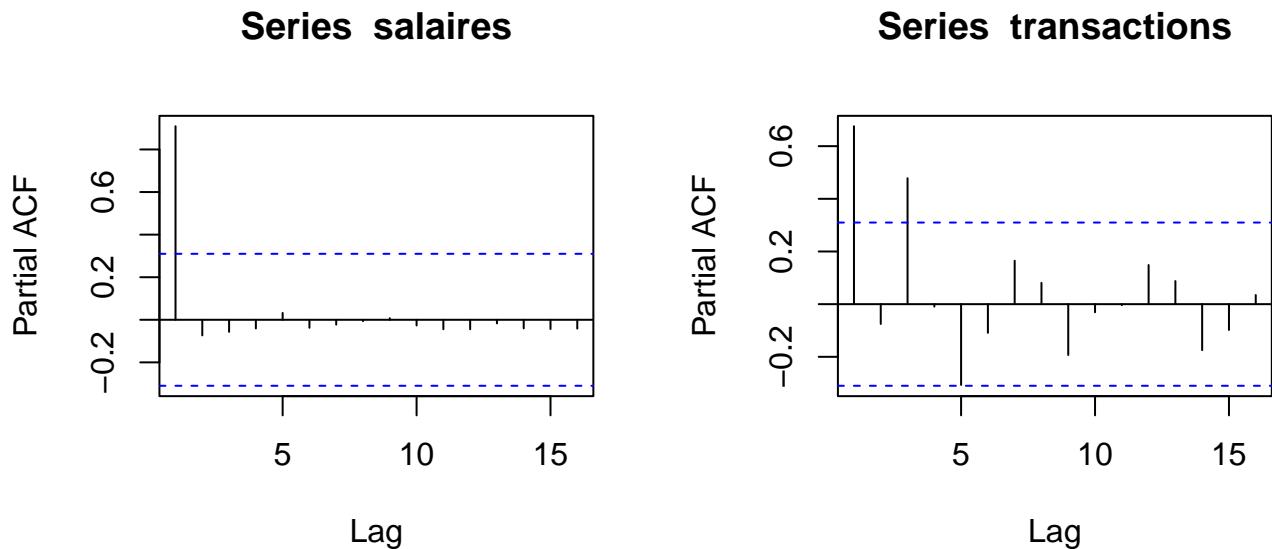
## 5.2 Analyse de l'auto-corrélation

Afin d'approfondir notre analyse temporelle, nous allons dans un premier temps examiner la structure de dépendance interne de chacune des deux séries à l'aide des fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF). Ces représentations graphiques permettent d'évaluer la persistance des effets dans le temps au sein d'une série donnée, en montrant comment chaque observation est corrélée à ses valeurs passées à différents retards (lags).

La courbe ACF mesure la corrélation entre une variable et ses valeurs décalées, tandis que la PACF mesure la corrélation entre une variable et ses valeurs décalées après avoir supprimé l'effet des autres décalages. Ces outils sont particulièrement utiles pour détecter la présence d'une tendance, d'un effet de saisonnalité.

Les graphiques des ACF et PACF des deux séries sont donc présentés ci-dessous afin d'évaluer visuellement leur comportement temporel et de détecter d'éventuelles structures d'autodépendance.





Pour les salaires, nous remarquons que les valeurs précédentes ont un impact important sur les valeurs actuelles, ce qui montre que l'évolution de cette variable se fait de manière progressive, avec des effets qui s'étalent dans le temps, comme attendu pour des salaires. Cela peut s'expliquer par des ajustements progressifs dus à des négociations salariales. Le graphique PACF montre surtout une influence marquée du premier retard, ce qui signifie que l'effet d'un changement est immédiat, et s'estompe ensuite.

Concernant les transactions, la dynamique semble un peu plus irrégulière. Nous observons une dépendance dans le temps, mais moins régulière que pour les salaires. Cela pourrait indiquer que les transactions réagissent plus fortement à des événements extérieurs comme les conditions de crédit par exemple. La PACF suggère que plusieurs retards peuvent jouer un rôle, pas seulement le premier, ce qui laisse penser que les effets mettent plus de temps à se diffuser.

En résumé, les salaires montrent une évolution assez régulière et progressive dans le temps, tandis que les transactions semblent réagir de manière un peu plus variable et moins prévisible.

### 5.3 Cointégration et modélisation

#### 5.3.1 Détection d'une cointégration

Nous avions vérifié que les séries étaient bien  $I(1)$  et désormais une question fondamentale se pose : évoluent-elles ensemble dans le temps malgré leur non-stationnarité individuelle ? Existe-t-il une

relation d'équilibre stable à long terme entre le nombre de transactions immobilières et les salaires mensuels de base ?

C'est précisément l'objectif d'un test de cointégration : détecter si deux séries non stationnaires sont liées par une combinaison linéaire stationnaire. Bien qu'elles puissent suivre des trajectoires instables prises séparément, il est possible qu'elles entretiennent un lien structurel qui les ramène l'une vers l'autre au fil du temps.

Pour répondre à cette problématique, nous réalisons un test de cointégration de Johansen (Johansen (1991)), particulièrement adapté à un cadre multivarié. Ce test repose sur une représentation en modèle VAR (vector autoregressive) des séries en différences premières et permet d'estimer le rang de cointégration, c'est-à-dire le nombre de relations cointégrantes entre les séries. Deux statistiques principales sont utilisées : la trace et la valeur propre maximale, chacune testant des hypothèses successives sur le nombre de relations de cointégration.

La réalisation de ce test constitue une étape essentielle dans notre démarche. En cas de cointégration, cela signifierait qu'un mécanisme d'ajustement à long terme lie nos deux séries, malgré des fluctuations de court terme. Cela justifierait alors l'estimation d'un modèle VECM (Vector Error Correction Model), capable de modéliser conjointement les dynamiques de court et long terme. En l'absence de cointégration, un simple VAR en différences pourrait être privilégié.

Nous allons donc maintenant appliquer le test de la trace de Johansen afin de détecter la présence ou non d'une relation de cointégration entre nos deux variables économiques. Nous choisissons un lag de 2 pour éviter la surparamétrisation du modèle. C'est un lag suffisant compte tenu du faible nombre d'observations.

Le test s'effectue en 2 étapes :

- $r \leq 1$  : L'hypothèse nulle est qu'il existe au plus une relation de cointégration.
- $r = 0$  : L'hypothèse nulle est qu'il n'existe aucune relation de cointégration.

Table 4: Test de cointégration de Johansen (test de trace au seuil de 5%)

Hypothèses	Test sans constante ni trend	Test avec constante	Test avec trend
$H_0 : r \leq 1$	0	4.18	4.27
$H_0 : r = 0$	15.68	27.17*	42.8*

Il est important d'identifier le terme déterministe dans la relation cointégrante. Les résultats dépendent du traitement de ce terme déterministe. Ainsi, chacun des 3 modèles a été testé pour la présence d'une relation de long terme entre les variables, le premier sans terme déterministe, le second avec un terme constant et le dernier avec la présence d'une trend dans la relation cointégrante.

- Test sans constante ni trend : D'après ce test, dans un premier temps, nous ne rejettions pas l'hypothèse nulle qu'il y ait au plus une relation mais ensuite, nous ne rejettions pas l'hypothèse qu'il n'y ait pas de relation. Ainsi, il n'existe pas de relation de cointégration entre les variables si nous considérons l'absence d'un terme déterministe.
- Test avec constante / Test avec trend : Ces 2 tests nous donnent les mêmes conclusions : il existe une relation de cointégration entre les 2 séries. En effet, dans un premier temps, nous ne rejettions pas l'hypothèse nulle qu'il y ait au plus une relation mais ensuite, nous rejettions l'hypothèse qu'il n'y ait pas de relation.

Désormais, nous devons choisir parmi ces trois modèles lequel représente au mieux la relation entre nos séries. Tout d'abord, nous savons que le modèle sans constante ni trend est très restrictif et souvent peu probable lorsque nous observons des données réelles d'après Johansen. Nous ne choisirons pas ce modèle qui ne parvient pas à capturer la relation entre nos variables car trop limité. Nous admettons donc la présence d'une relation cointégrante entre les deux séries, les deux autres modèles arrivant à déterminer une dynamique de long terme, nous devons choisir le plus pertinent au vu de nos séries. Le modèle avec trend serait justifiable économiquement. En effet, puisque nous observons une trend déterministe dans les séries en niveau des salaires et du nombre de transactions, il serait logique de penser qu'une trend déterministe pourrait être présente dans

la relation qui les lie. Cependant, nous pouvons évoquer le Pantula Principle (Pantula, Gonzalez-Farias, and Fuller (1994)), comme appliqué dans cet article Sinha and Mbulawa (2023) . C'est une démarche académique de spécification de modèle. Cette approche propose de tester la présence de relation de cointégration dans le modèle le plus général : celui avec trend. Si ce test s'avère positif, nous testons ensuite le modèle avec constante seulement. De la même manière, si une relation est captée, nous testons le modèle sans constante ni trend. La finalité du Pantula Principle est de s'arrêter au modèle le plus simple qui arrive à détecter une relation cointégrante.

D'après Pantula, le modèle avec constante est le modèle le plus simple qui suffit pour capter l'équilibre de long terme. Finalement, au vu des éléments, nous décidons de retenir le modèle avec constante dans l'équilibre, plus stable et moins complexe qu'un modèle avec trend dans l'équilibre.

### 5.3.2 Modélisation VECM

Il est donc maintenant nécessaire de modéliser et d'estimer un modèle VECM, compte tenu de la relation de cointégration. Ce modèle s'écrit :

$$\Delta Y_t = \alpha \beta' Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \dots + \Gamma_{p-1} \Delta Y_{t-p+1} + \Theta D_t + \epsilon_t$$

Avec :

- $\Delta Y_t$  le vecteur des premières différences des variables au temps  $t$ .
- $\alpha$  la matrice d'ajustement.
- $\beta'$  la transposée du vecteur de cointégration.
- $\beta' Y_{t-1}$  le terme de correction d'erreur, qui mesure le déséquilibre par rapport à la relation de long terme au temps  $t - 1$ .
- $\Gamma_i$  les matrices de coefficients des différences retardées des variables.
- $\Theta D_t$  qui représente les termes déterministes (ici la constante).
- $\epsilon_t$  le vecteur des termes d'erreur.

Dans notre cas, nous pouvons identifier les éléments de cette équation.

Nous estimons les coefficients de la relation de cointégration (le vecteur  $\beta$ ). Ils indiquent la combinaison linéaire des variables qui est stationnaire à long terme, normalisée à 1. Cette relation s'écrit :

$$\beta = \begin{pmatrix} 1.00 \\ -0.4697 \\ -57.3551 \end{pmatrix}$$

Nous pouvons écrire l'équation de cointégration telle que :

$$1.00 \times \text{salaires} - 0.4697 \times \text{transactions} = 57.3551 \times \text{const}$$

Puis :

$$\text{salaires} = 0.4697 \times \text{transactions} + 57.3551$$

Ainsi, une augmentation des transactions immobilières en base 100 de 1 unité est associée à une augmentation de 0.4697 unité des salaires mensuels de base en base 100 à long terme, en tenant compte de la constante.

Le modèle VECM estimé se présente comme :

$$\begin{aligned} \Delta\text{salaires}_t &= -0.0158^* \cdot \text{ECT} + 0.1981 \cdot \Delta\text{salaires}_{t-1} \\ &\quad - 0.0129^* \cdot \Delta\text{transactions}_{t-1} + 0.3459 \cdot \Delta\text{salaires}_{t-2} - 0.000043 \cdot \Delta\text{transactions}_{t-2} \end{aligned}$$

$$\begin{aligned} \Delta\text{transactions}_t &= -0.1556 \cdot \text{ECT} + 0.5289 \cdot \Delta\text{salaires}_{t-1} \\ &\quad - 0.1706 \cdot \Delta\text{transactions}_{t-1} - 6.7337 \cdot \Delta\text{salaires}_{t-2} - 0.6846^* \cdot \Delta\text{transactions}_{t-2} \end{aligned}$$

Avec \* qui distingue les coefficients statistiquement significatifs au seuil de 5%.

L'analyse du modèle VECM met en évidence plusieurs éléments d'interprétation statistiquement significatifs et d'autres non.

La matrice des poids, appelée matrice d'ajustement (la matrice  $\alpha$ ), qui représente les forces de rappel et mesure la vitesse à laquelle chaque variable s'ajuste pour revenir à la relation d'équilibre de long terme, s'écrit :

$$\alpha = \begin{pmatrix} -0.0158 \\ -0.1556 \end{pmatrix}$$

Seul le coefficient associé à la variable des salaires est significatif à 5 %.

Le coefficient de -0.0158 signifie qu'en cas de déséquilibre par rapport à la relation de long terme, les salaires tendent à s'ajuster lentement pour revenir à l'équilibre. Le déséquilibre suite à une déviation de la relation de long terme est corrigé d'environ 1.58% par période. Le temps nécessaire pour corriger 50% de cet écart peut être calculé par cette formule :  $(1 - 0.0158)^t = 0.5$  donc  $t \approx 43.6$ . Il faut environ 44 périodes pour que les salaires corrige 50% de l'écart, ce qui est particulièrement lent.

A l'inverse, bien que le coefficient d'ajustement sur les transactions soit plus fort (15.56%), il n'est pas significatif, ce qui empêche de conclure qu'il revient activement vers l'équilibre. S'il avait été significatif, nous aurions pu calculer le temps nécessaire pour corriger 50% de cet écart :  $(1 - 0.1556)^t = 0.5$  donc  $t \approx 4.1$ . Il aurait fallu environ 4 périodes aux transactions pour corriger 50% de l'écart. Cela aurait été un résultat cohérent. Cependant, ce n'est pas statistiquement significatif et donc pas fiable donc nous ne pouvons pas prendre en compte ce résultat.

Dans la première équation, celle des salaires, seuls le terme de correction d'erreur et les transactions en  $t - 1$  sont significatifs à 5%. Le coefficient de -0.0129 indique que les variations passées du nombre de transactions ont un effet légèrement négatif sur les variations actuelles des salaires. Ce résultat est économiquement surprenant : nous nous attendions à ce que les transactions n'influencent pas les salaires, simplement car nous pensions que les salaires étaient plus exogènes dans cette relation. Mais en plus, le fait que ce coefficient soit négatif ne semble pas cohérent avec les théories économiques, un marché immobilier dynamique s'accorderait logiquement avec une hausse des salaires. Le fait que ce soit l'inverse est peut-être dû à des facteurs spécifiques au contexte économique ou par l'influence d'autres variables non prises en compte dans le modèle.

Dans la seconde équation, celle des transactions, seul le coefficient des transactions en  $t - 1$  ressort significatif. Le coefficient négatif de -0.6846 suggère une forme d'effet de correction : une hausse des transactions deux périodes plus tôt est suivie d'une baisse. Ce type de dynamique sur le marché de l'immobilier est cohérent. En effet, pour rappel, lorsque nous avions représenté la série en niveau, nous observions une grande volatilité des observations qui oscillaient sans cesse. Aucun effet significatif des salaires passés sur les transactions n'a été détecté, ce qui affaiblit l'hypothèse selon laquelle

les revenus influencerait directement et immédiatement le nombre d'échanges immobiliers à court termes. Cette absence de lien significatif peut refléter une complexité plus forte du comportement d'achat, influencé par le crédit, la confiance des ménages ou la politique publique.

#### 5.4 Vérification des résidus du modèle

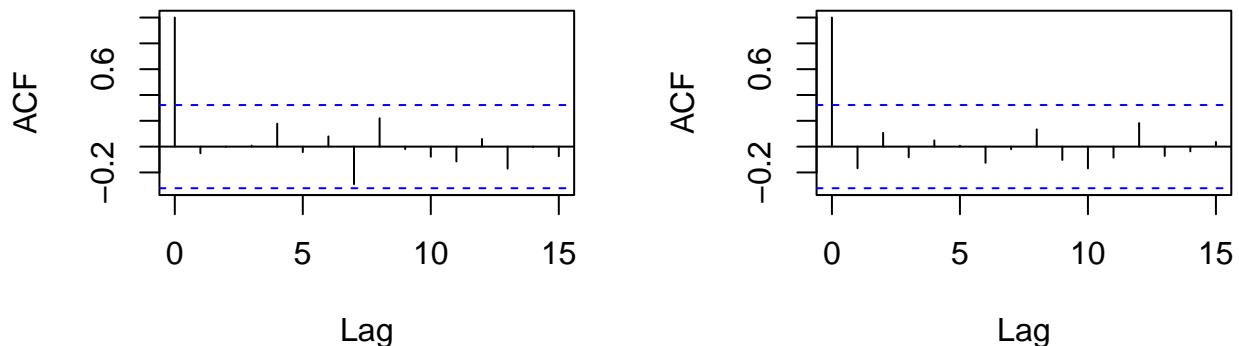
Afin de garantir la validité des résultats précédents, il est nécessaire que les hypothèses classiques soient respectées. Ainsi, nous avons réalisé une série de tests sur les résidus du modèle.

Autocorrélation des résidus :

Les résidus doivent être non autocorrélés pour respecter les hypothèses.

Nous pouvons représenter les fonctions d'autocorrélation des résidus de chaque variable.

#### ACF des résidus – variable salaire ACF des résidus – variable transacti



Graphiquement, il semble évident que les résidus ne sont pas corrélés. Afin de confirmer cette intuition, le test de Portmanteau (Ljung-Box) a été réalisé. Ses hypothèses sont :

$$\begin{cases} H_0 : \text{Les résidus ne sont pas autocorrélés} \\ H_1 : \text{Les résidus sont autocorrélés} \end{cases}$$

Table 5: Résultat du test d'autocorrélation des résidus

	Test	p.value
Chi-squared	Portmanteau (autocorrélation)	0.3487

Nous concluons donc que le test ne détecte pas d'autocorrélation significative dans les résidus du modèle. Cela indique que la dynamique temporelle du VECM capture correctement les dépendances présentes dans les données.

#### Hétéroscedasticité des résidus :

Les résidus doivent être homoscédastiques pour respecter les hypothèses.

Nous avons réalisé le test ARCH qui va tester la présence d'homoscédasticité conditionnelle dans le modèle. Ses hypothèses sont :

$$\begin{cases} H_0 : \text{Pas d'effet ARCH} \rightarrow \text{variance constante des résidus} \\ H_1 : \text{Effet ARCH} \rightarrow \text{variance des résidus dépendante du passé} \end{cases}$$

Table 6: Résultat du test d'hétéroscedasticité (ARCH)

	Test	p.value
Chi-squared	ARCH (hétéroscedasticité)	0.6582

La variance des résidus est bien constante.

#### Normalité des résidus :

Enfin, il est nécessaire que les résidus de notre modèle soient distribués selon une loi normale pour respecter les hypothèses.

Nous avons effectué un test de normalité dont voici les hypothèses :

$$\begin{cases} H_0 : \text{Les résidus suivent une distribution normale} \\ H_1 : \text{Les résidus ne suivent pas une distribution normale} \end{cases}$$

Table 7: Résultat du test de normalité des résidus

Test	p.value
Normalité (Jarque-Bera multivarié)	4.03e-14

Ce test conclut à la non-normalité des résidus. Cela pose un problème car tous nos résultats précédents sont désormais remis en question.

Ainsi, toutes les hypothèses essentielles sur les résidus ne sont pas respectées. Nous ne pouvons pas garantir la validité du modèle VECM estimé.

L'analyse menée dans cette partie s'inscrit dans une démarche de compréhension des dynamiques économiques sur le long et le court terme. L'approche par séries temporelles, et en particulier la modélisation VECM, nous a permis d'explorer la manière dont nos variables évoluent dans le temps et s'influencent mutuellement, avec des retards et des ajustements.

Ce type se présentait comme particulièrement pertinent dans le cadre de variables comme les salaires et le nombre de transactions immobilières, qui sont toutes deux soumises à des évolutions progressives, à des cycles économiques.

Nous soupçonnions que les salaires et l'activité immobilière, représentée ici par le nombre de logements changeant de propriétaire, présentent une tendance à évoluer ensemble. Ce lien aurait reflété plusieurs mécanismes : par exemple, une hausse des salaires peut soutenir la demande de logements, tandis qu'un marché immobilier actif peut traduire une situation économique favorable susceptible d'augmenter les salaires.

Malheureusement, nous ne pouvons tirer de conclusions quant à cette modélisation temporelle. Les hypothèses essentielles du modèle n'étant pas respectées, nous ne pouvons garantir la fiabilité de ce VECM. Il est important de garder à l'esprit que nos résultats restent influencés par la taille réduite de l'échantillon (40 observations), l'absence de variables explicatives complémentaires (taux d'intérêt, inflation, chômage), et une éventuelle hétérogénéité régionale non prise en compte. Des analyses plus poussées pourraient intégrer de nouvelles variables dans le modèle, pour mieux capter les dynamiques des variables. En résumé, cette analyse non concluante a présenté ses limites et n'a pas permis d'établir de lien temporel significatif entre les salaires et le nombre de mutations immobilières. Leur relation, qui semble complexe, nécessiterait d'être considérée autrement.

## 6 Conclusion

L'ensemble de nos analyses menées au cours de ce travail ont permis de confirmer la pertinence du recours à l'économétrie spatiale pour expliquer les disparités des salaires à l'échelle des zones d'emplois de la France. Lorsque nous nous sommes intéressés au lien entre le nombre de logements changeant de propriétaires et le salaire net horaire, notre étude a montré des résultats nuancés. L'utilisation du modèle spatial SDM a intégré des effets locaux mais également ceux des zones voisines, ce qui a permis de capturer non seulement l'effet direct d'un changement dans une zone mais également l'effet indirect des dynamiques des zones limitrophes. Nos résultats montrent une forte autocorrélation spatiale justifiant le non recours à la modélisation OLS. Les effets spatiaux estimés attestent que les niveaux de salaires ne peuvent pas être considérés pleinement.

Par ailleurs, l'étude des séries temporelles et du panel dans ce travail nous invite à la prudence : le nombre limité de périodes disponibles ne permet pas d'identifier des dynamiques longues ou robustes et pourrait affecter la stabilité des coefficients estimés. Ces limites méthodologiques appellent ainsi à des prolongements d'analyses futurs notamment par l'intégration de données à plus haute fréquence permettant de mieux capter les évolutions structurelle à long terme. Les résultats indiquent que si les mutations immobilières sont faiblement explicatives du salaire local, une fois d'autres variables de contrôles incluses, elles présentent néanmoins un effet indirect négatif qui peut être illustré par une forme de concurrence territoriale. À l'inverse, la densité d'emploi apparaît comme un facteur déterminant de structures salariales, quant au taux de chômage son effet direct est logiquement négatif mais son effet indirect suggère un phénomène de rareté et d'attractivité locale.

Au-delà des apports empiriques apportés, cette étude s'interroge sur des dimensions fondamentales de la géographie économique telle que l'attractivité résidentielle comme vecteur de développement économique ou la capacité des territoires à générer et redistribuer des richesses. Ces résultats ouvrent des pistes de réflexion pour les politiques publiques visant à réduire les inégalités territoriales. Des actions ciblées pour développer l'emploi qualifié et diversifier l'économie dans les zones défavorisées pourraient contribuer à réduire les écarts salariaux. En outre, notre étude souligne la complexité des dynamiques spatiales qui façonnent les rémunérations salariales en France. Elle invite à une réflexion nuancée sur les politiques de développement territorial, prenant en compte les spécificités locales et les interdépendances régionales.

## 7 Bibliographie

### Livres articles :

- Dabet, Gaelle, and Jean-Michel Floch. n.d. "Direction de La Diffusion Et de l'action Regionale."
- De Hoyos, Rafael E, and Vasilis Sarafidis. 2006. "Testing for Cross-Sectional Dependence in Panel-Data Models." *The Stata Journal* 6 (4): 482–96.
- Hoechle, Daniel. 2007. "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence." *The Stata Journal* 7 (3): 281–312.
- Johansen, Søren. 1991. "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models." *Econometrica: Journal of the Econometric Society*, 1551–80.
- Moran, Patrick AP. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37 (1/2): 17–23.
- Pantula, Sastry G, Graciela Gonzalez-Farias, and Wayne A Fuller. 1994. "A Comparison of Unit-Root Test Criteria." *Journal of Business & Economic Statistics* 12 (4): 449–59.
- Sinha, Narain, and Strike Mbulawa. 2023. "Government Expenditure on Health and Economic Growth in Botswana: Testing for Cointegration and Specification of Deterministic Components Using the Pantula Principle." *International Journal of Research in Business and Social Science* 12 (2): 204–16.

Cours Panel Master ESA M1 C.Rault (2024)

Cours Series temporelles Master ESA M1 G. De Truchis (2024)

### Sitographie :

Voici les liens de nos différentes sources de données :

Indicateur Immobilier Indice salaires Carte ZE SNHM 2022 Emploi localisé Chomage ZE IPC

L'ensemble de nos données proviennent du site de l'INSEE et des données fournies.

Dimension temporelle :

- <https://stats.stackexchange.com/questions/97783/understanding-vec2var-conversion-in-r>
- <https://www.rdocumentation.org/>
- <https://www.ibm.com/fr-fr/think/topics/autocorrelation#:~:text=La%20fonction%20d'autocorr%C3%A9ation>

- <https://www.r-bloggers.com/2021/11/granger-causality-test-in-r-with-example/>
- <https://www.r-econometrics.com/timeseries/irf/>
- <https://www.quantstart.com/articles/Johansen-Test-for-Cointegrating-Time-Series-Analysis-in-R/>
- <https://www.aptech.com/blog/how-to-interpret-cointegration-test-results/>
- <https://bookdown.org/jarneric/econometrics/9.2-cointegration.html>

## 8 Annexe

Table 8: Focus sur les Salaires Nets Horaires Moyens par Sexe, CSP et Tranche d'Age

SNHM	SNHMC	SNHMP	SNHME	SNHMO	SNHMF
Min. :14.28	Min. :25.54	Min. :14.59	Min. :10.56	Min. :11.03	Min. :12.66
1st Qu.:14.60	1st Qu.:26.04	1st Qu.:14.75	1st Qu.:10.70	1st Qu.:11.08	1st Qu.:13.01
Median :15.10	Median :26.92	Median :15.06	Median :10.84	Median :11.31	Median :13.61
Mean :15.36	Mean :27.05	Mean :15.31	Mean :11.09	Mean :11.51	Mean :13.84
3rd Qu.:16.19	Qu.:28.01	Qu.:15.87	Qu.:11.52	Qu.:11.93	Qu.:14.70
Max. :17.02	Max. :29.09	Max. :16.58	Max. :12.12	Max. :12.50	Max. :15.60

SNHMFC	SNHMFP	SNHMFE	SNHMFO	SNHMH	SNHMHC
Min. :21.94	Min. :13.51	Min. :10.33	Min. : 9.632	Min. :15.39	Min. :27.36
1st Qu.:22.60	1st Qu.:13.65	1st Qu.:10.46	1st Qu.: 9.790	1st Qu.:15.71	1st Qu.:27.87
Median :23.55	:14.06	:10.65	:10.062	:16.15	:28.78
Mean :23.74	Mean :14.25	Mean :10.89	Mean :10.207	Mean :16.43	Mean :28.86
3rd Qu.:24.93	Qu.:14.83	Qu.:11.36	Qu.:10.617	Qu.:17.25	Qu.:29.79
Max. :26.09	Max. :15.54	Max. :12.00	Max. :11.214	Max. :18.02	Max. :30.88

SNHMHP	SNHMHE	SNHMHO	SNHM18	SNHM26	SNHM50
Min. :15.39	Min. :11.02	Min. :11.28	Min. : 9.610	Min. :14.15	Min. :16.84
1st Qu.:15.60	1st Qu.:11.16	1st Qu.:11.33	1st Qu.: 9.723	1st Qu.:14.38	1st Qu.:17.28
Median :15.89	Median :11.30	Median :11.57	Median : 9.933	Median :14.87	Median :17.85

SNHMHP	SNHMHE	SNHMHO	SNHM18	SNHM26	SNHM50
Mean :16.17	Mean :11.50	Mean :11.78	Mean :10.215	Mean :15.16	Mean :18.04
3rd	3rd	3rd	3rd	3rd	3rd
Qu.:16.76	Qu.:11.87	Qu.:12.21	Qu.:10.675	Qu.:15.97	Qu.:18.88
Max. :17.50	Max. :12.38	Max. :12.78	Max. :11.434	Max. :16.85	Max. :19.68

SNHMF18	SNHMF26	SNHMF50	SNHMH18	SNHMH26	SNHMH50
Min. : 9.251	Min. :12.79	Min. :13.96	Min. : 9.891	Min. :15.07	Min. :18.81
1st Qu.: 9.378	1st :13.09	1st :14.43	1st Qu.: 9.990	1st :15.27	1st :19.27
Median : 9.606	Median :13.71	Median :15.08	Median :10.186	Median :15.66	Median :19.80
Mean : 9.881	Mean :13.95	Mean :15.27	Mean :10.471	Mean :15.99	Mean :19.99
3rd Qu.:10.362	3rd Qu.:14.82	3rd Qu.:16.18	3rd Qu.:10.914	3rd Qu.:16.77	3rd Qu.:20.79
Max. :11.105	Max. :15.76	Max. :17.06	Max. :11.677	Max. :17.61	Max. :21.56

## 8.1 Dimension temporelle

Table 12: P-values des tests PP et ADF

Série	p value PP	p value ADF
Salaires (niveau)	0.9900	0.9900
Transactions (niveau)	0.2338	0.9900
Salaires (diff)	0.0100	0.8673
Transactions (diff)	0.0100	0.1238

## 8.2 Modèle Spatial

```

Moran I test under randomisation

data: zones_log$log_salaire
weights: listw_log

Moran I statistic standard deviate = 13.441, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic      Expectation      Variance
0.500032969      -0.003610108      0.001404137

```

Figure 28: Annexe 1

```

Moran I test under randomisation

data: res_ols
weights: listw_ols

Moran I statistic standard deviate = 6.49, p-value = 4.292e-11
alternative hypothesis: greater
sample estimates:
Moran I statistic      Expectation      Variance
0.241853780      -0.003610108      0.001430491

```

Figure 29: Annexe 2

Model	df	AIC	logLik	Test L.Ratio	p-value
sar_log	1	4	-763.67	385.83	1
sdm_log	2	5	-778.17	394.08	2 16.502 4.861e-05

Figure 30: Annexe 3

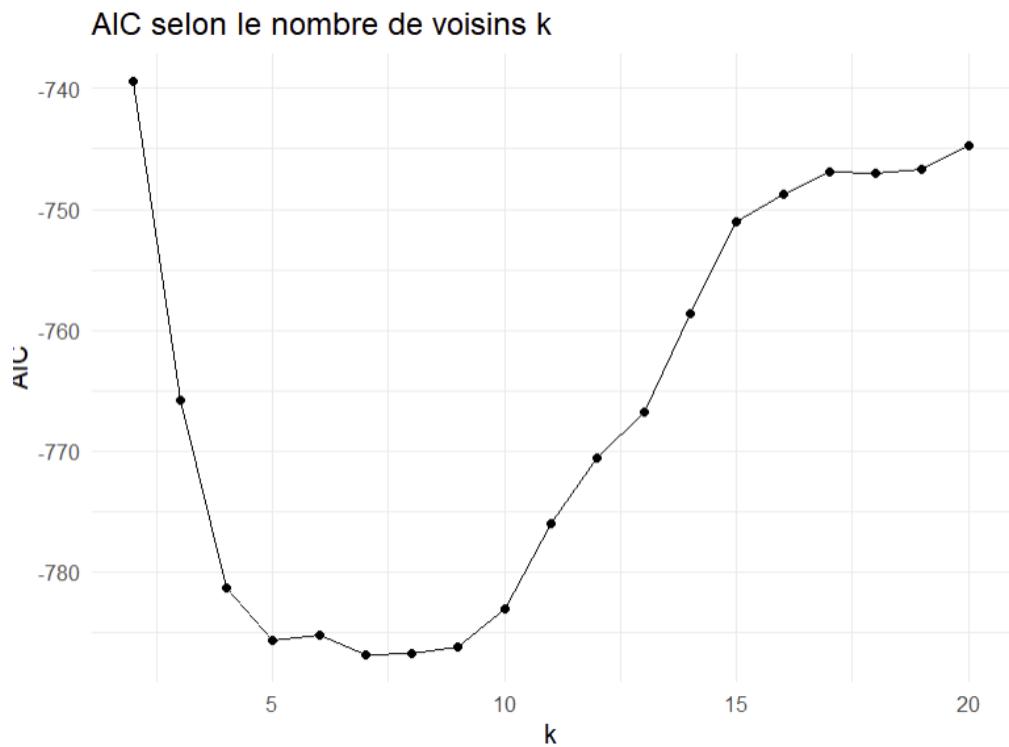


Figure 31: Annexe 4

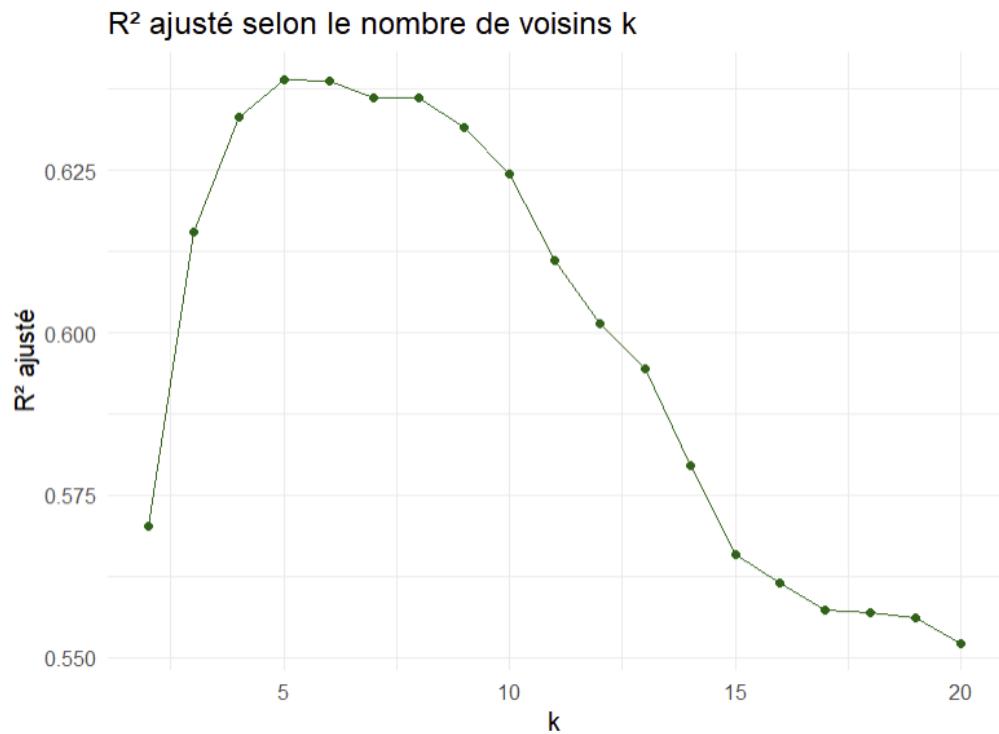


Figure 32: Annexe 5

Table 5: Test de Shapiro-Wilk pour la normalité des résidus du modèle SDM (KNN = 5)

Modèle	Statistique W	p-value	Conclusion
SDM (KNN = 5)	0.9629	0	Non normalité détectée

Figure 33: Annexe 6

Table 1: Corrélation entre les variables explicatives

	log_salaire	log_emploi	TauxChomage2022
log_salaire	1.000	0.608	-0.082
log_emploi	0.608	1.000	0.104
TauxChomage2022	-0.082	0.104	1.000

Figure 34: Annexe 7

Résultats du modèle SDM (log salaire ~ variables sociales et économiques)

	Variable	Estimate	Std. Error	z-value	Pr(> z )
(Intercept)	(Intercept)	0.982	0.209	4.698	0.000
Part_cadres_chefs	Part_cadres_chefs	0.011	0.001	11.703	0.000
Part_employes_ouvriers	Part_employes_ouvriers	-0.003	0.001	-3.599	0.000
log_transaction	log_transaction	-0.005	0.003	-2.085	0.037
TauxChomage2022	TauxChomage2022	-0.001	0.001	-0.537	0.591
lag.Part_cadres_chefs	lag.Part_cadres_chefs	-0.008	0.002	-3.741	0.000
lag.Part_employes_ouvriers	lag.Part_employes_ouvriers	0.000	0.002	0.231	0.817
lag.log_transaction	lag.log_transaction	0.007	0.005	1.426	0.154
lag.TauxChomage2022	lag.TauxChomage2022	0.000	0.002	-0.310	0.756

Figure 35: Annexe 8

	VIF
log(TauxChomage2022)	1.010
log(SNHM22)	1.297
Emploi2022	1.293

Figure 36: Annexe 9

Table 2: Test d'hétérosécédasticité modèle Modèle enrichi

	Variable	Coefficient	Erreur.Std.	t.value	p.value
(Intercept)	(Intercept)	-0.00222	0.00327	-0.68	0.498
log_transactions	log_transactions	0.00119	0.00081	1.47	0.144
log_emploi	log_emploi	-0.00038	0.00071	-0.53	0.597
chomage	chomage	-0.00016	0.00015	-1.04	0.299

Figure 37: Annexe 10

Table 2: Test d'hétérosécédasticité : Modèle simple)

	Variable	Coefficient	Erreur.Std.	t.value	p.value
(Intercept)	(Intercept)	-0.00574	0.00374	-1.53	0.1260
log_transactions	log_transactions	0.00101	0.00047	2.17	0.0311

Figure 38: Annexe 11

### 8.3 Panel

Comme évoqué dans la partie Panel par ZE, nous avons préalablement fait une analyse sur les départements au vu du redécoupage des zones d'emploi entre les 2 décennies qui semble assez drastique dans sa forme. Nous supposons que ces différences auraient un impact trop fort sur nos résultats.

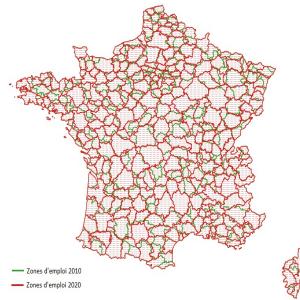


Figure 39: Redécoupage des ZE

#### Présentation du panel

Avant de passer à une quelconque analyse, nous proposons de présenter le panel utilisé par la suite. Il est de dimensions 837 par 26. Au sein de celui-ci nous pouvons trouver les variables *année* et *departements* ainsi que les *transactions* et *salaires* et un nombre important de variables de contrôle que nous utiliserons dans un second temps.

##### 8.3.1 Variables

Variables expliquées :

Les variables que nous allons modéliser sont les transactions (changement de propriétaire) et le salaire net horaire moyen. En 2022, leurs valeurs moyennes étaient réparties ainsi au sein des départements :

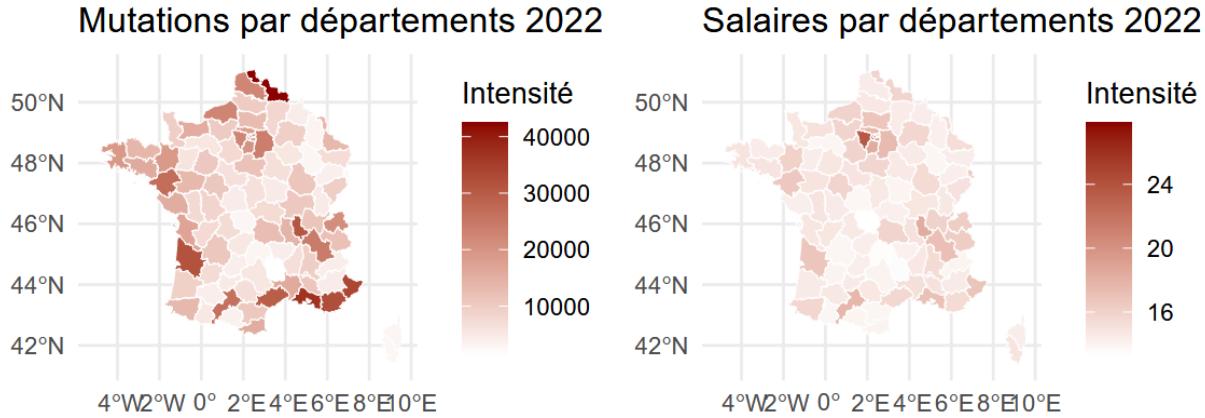


Figure 40: Carte

Cette répartition est globalement similaire de 2014 à 2022 avec quelques légères évolutions. Nous remarquons qu'entre les régions dites rurales et non rurales, nous observons de grandes disparités. Il serait intéressant de prendre en compte ceci par la suite. Les salaires et transactions sont vraiment bas dans la diagonale du vide et plus élevés dans les régions actives et bassins d'emploi comme nous pourrions nous en douter.

Regardons cela de plus près en séparant les départements ruraux et non-ruraux :

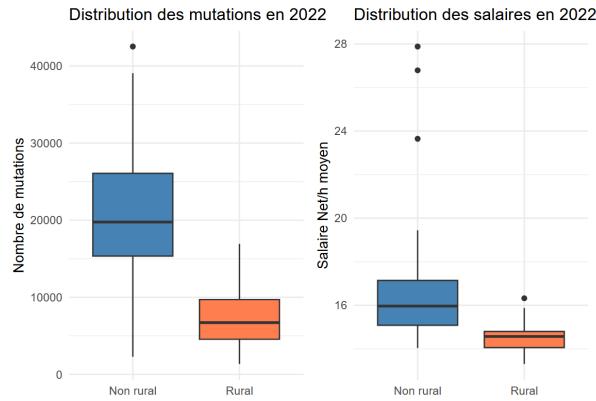


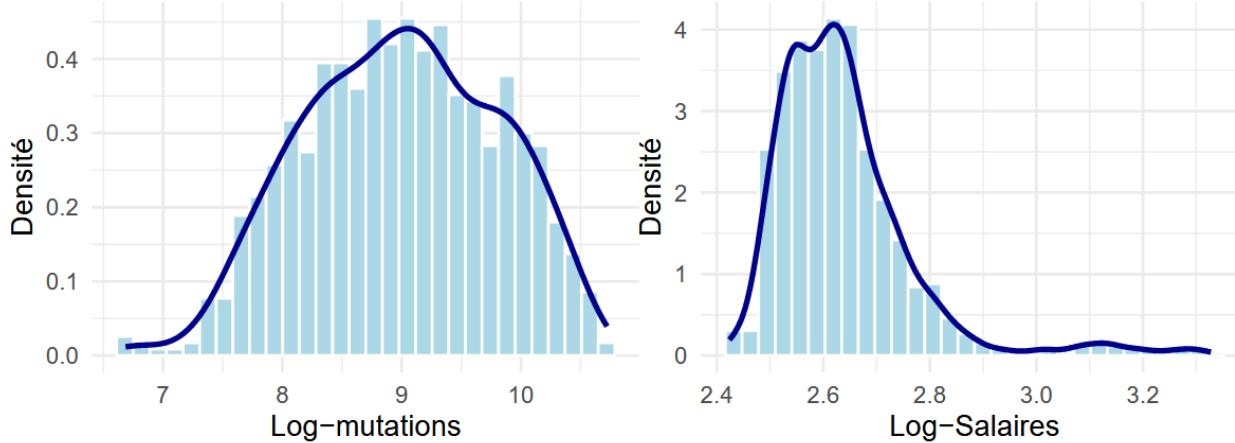
Figure 41: boxplot

Ici, il est vraiment clair qu'il y a une différence significative entre les régions rurales et non-rurales, nous la prendrons en compte dans nos analyses futures, notamment dans la robustesse de nos coefficients. Le salaire net horaire moyen médian a presque 1.5 euros d'écart ce qui est énorme

mais peut s'expliquer par le coût de la vie dans ces différentes régions.

Compte tenu de ces grandes disparités et pour améliorer nos futurs modèles, nous préférerons les variables en logarithme. Les interprétations seront alors des élasticités. Nous réduisons ainsi l'écart entre les valeurs maximales et minimales pour diminuer les variances et rendre plus symétriques nos variables.

## Distributions



Les séries semblent assez symétriques désormais si ce n'est les salaires de la région parisienne.

Variables contrôle :

Nous avons opté, et détaillerons ces choix par la suite, pour la part d'emploi au sein d'un département dans différents secteurs ainsi que le chômage et la part de jeunes (25-39ans) et vieux (65+ ans) travailleurs. Nous introduisons par la même occasion une variable indiquant le nombre d'emplois et la population au sein du département.

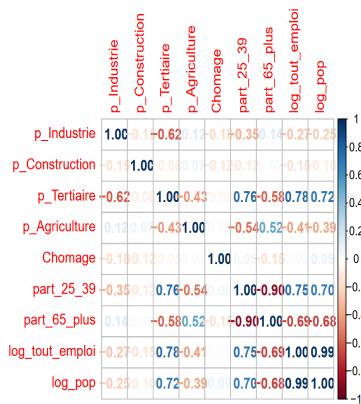


Figure 42: Corrélation

### 8.3.2 L'impact des mutations sur les salaires

Nous allons nous intéresser à l'impact des transactions sur les salaires. Le marché résidentiel influence t-il les salaires locaux ? Celui-ci est-il quantifiable et significatif ? Nous tenterons de répondre à ces questions via l'approche panel. Bien qu'il n'y ait pas de littérature sur le sujet, nous pouvons émettre quelques hypothèses sur les résultats attendus. Nous nous attendons à 2 mécanismes :

Un premier effet bénéfique que nous appelons d'“*accommodation*”. La quantité de transactions indiquerait une certaine fluidité du marché résidentiel facilitant donc la mobilité des travailleurs et permettant ainsi une certaine accommodation entre leurs compétences et emplois. Cela augmenterait le coût de rester dans un emploi non optimal pour le travailleur l'incitant donc à mieux s'apparier et ayant pour conséquences d'augmenter les salaires. Un coefficient positif serait alors attendu.

Le deuxième mécanisme, celui-ci plutôt négatif, est celui de la *pression des offres*. Si au sein de notre territoire nous observons une grande affluence au sein des transactions (exode, saisonnalité, ...), alors, l'offre de travail sera plus grande qu'auparavant ce qui va toute chose égale par ailleurs baisser les salaires moyens. Un coefficient négatif serait alors attendu.

L'enjeu ici sera de révéler lequel de ces deux mécanismes domine afin d'établir l'impact que les transactions ont sur les salaires. Enfin, nous pourrions aussi nous attendre à des effets de composition sectorielle, à de l'asymétrie générationnelle, etc... qui seront nos variables de contrôle par la suite afin de vérifier la robustesse de nos résultats.

**8.3.2.1 Modèle simple** Préliminairement à l'introduction d'une multitude de variables de contrôle démographiques, sectorielles et conjoncturelles dans les prochaines parties, il est important d'observer la relation “nue” ou “brute” entre nos 2 variables d'intérêt. Comme énoncé précédemment, nous prendrons le logarithme du salaire net horaire moyen ainsi que le logarithme du nombre de transactions.

Cette étape est cruciale pour les raisons suivantes, nous allons obtenir un coefficient qui regroupera l'effet des transactions mais aussi de potentielles autres variables l'influencant. Grâce à ceci, nous pourrons voir par la suite l'évolution de ce coefficient après l'ajout de variables de contrôle. Un autre point est celui des tests, avec ce modèle nous pourrons établir de potentielles incohérences.

Au vu de nos analyses sur les variables, nous construirons un modèle global et 2 autres modèles

portant sur la ruralité des départements. Aussi, nos modélisations futures porteront sur un modèle *Within* et cela n'entravera pas nos analyses puisque nous avons bien de la variance temporelle dans nos séries (ANNEXE) garantissant la non nullité de nos variables après *Within*. Par ailleurs, nous observons une corrélation moyenne de 0.6 entre nos 2 séries en moyenne au cours des années.

Notre modèle à effet individuel sera le suivant pour un département  $i$  à l'année  $t$  :

$$s_{i,t} = a + \beta m_{i,t} + \alpha_i + \delta_t + \epsilon_{i,t} \quad (1)$$

Avec :

- $s_{i,t}$  les log-salaires net horaire moyen
- $m_{i,t}$  les log-mutations
- $\alpha_i$  l'effet individuel
- $\delta_t$  l'effet temporel
- $a$  la constante
- $\epsilon_{i,t}$  le terme d'erreur

Ici, nous émettons l'hypothèse qu'il y a de l'effet individuel pour plusieurs raisons. Il existe un nombre important de variables difficilement observables influençant un département à avoir plus ou moins de transactions ou changement de salaire. Ces variables vont bouger lentement dans le temps voir pas, sauf évènement exceptionnel. Ainsi, il faut inclure de l'effet individuel (fixe nous le verrons par la suite) pour capter la véritable relation entre les transactions et salaires car sinon le modèle capterait simplement l'attractivité du département donc nous allons supprimer, via l'introduction de l'effet fixe, l'hétérogénéité permanente. Cela est d'une importance primordiale pour capter l'essence de ce que nous recherchons, l'impact des mutations sur les salaires.

L'effet temporel est aussi primordial pour capter les chocs macro-économiques communs aux départements comme par exemple le Covid. De plus  $\epsilon_{i,t}$  est non corrélé avec les effets et les variables explicatives. Cette stratégie garantit que le coefficient  $\beta$  repose exclusivement sur les variations intra-département qui demeurent, après extraction, des tendances nationales.

**8.3.2.1.1 Identification du modèle** Nous ne pouvons pas encore faire d'hypothèses sur le terme  $\alpha_i$ , bien que nous soupçonnions un modèle à effets fixes. Pour cela il faut le vérifier statis-

tiquement, notamment à l'aide de tests de spécification. Il existe 2 types de test utilisés en panel : le test de Fisher, pour vérifier s'il existe bien de l'effet individuel, et le test d'Hausman et celui de Mundlak pour vérifier si l'effet individuel est corrélé aux variables explicatives. En fonction de nos résultats, nous pourrons dire si nous avons un modèle à erreur simple, un REM (Random effect model) ou FEM (Fixed effect model) et donc choisir le bon estimateur (MCO, GLS, Within).

Test de Fisher :

Nos hypothèses sont les suivantes :

$$\begin{cases} H_0 : \sigma_\alpha^2 = 0 & \text{Il n'y a pas d'effet individuel} \\ H_1 : \sigma_\alpha^2 > 0 & \text{Il y a de l'effet individuel} \end{cases}$$

Le test se base sur les résidus des modèles estimés sous l'estimateur *between* et *within*. (Plus d'info voir le Cours Rault) La statistique de test suit une loi de Fisher :

$$F = \frac{\hat{\sigma}_W^2}{\hat{\sigma}_B^2} \rightsquigarrow F(N - p, NT - N - p + 1)$$

Avec  $N$  le nombre d'individus,  $T$  le nombre de périodes et  $p$  le nombre de variables.

En appliquant ce test à nos données, nous pouvons déterminer s'il y a de l'effet temporel et de l'effet individuel.

Voici nos résultats :

Table 13: Test F de Fisher

Stat.F	P.Value
85.735	0
17.890	0

Nous remarquons que pour nos 2 tests, nous rejetons l'hypothèse nulle  $H_0$ . Ainsi, nous concluons à la présence d'effet individuel et temporel dans notre modèle comme nous nous y attendions.

Test de Hausman :

Notre modèle est donc soit un modèle REM soit un modèle FEM. Pour arbitrer, nous devons effectuer le test de Hausman. Le test de Hausman se base sur une comparaison directe des estimateurs. Ce test nous permet de savoir si nos effets individuels sont corrélés avec nos variables explicatives.

$$\begin{cases} H_0 : \text{Plim}_{\bar{NT}} \frac{1}{NT} X' \alpha = 0 & \text{Effet individuel corrélé aux explicatives} \\ H_1 : \text{Plim}_{\bar{NT}} \frac{1}{NT} X' \alpha \neq 0 & \text{Non Corrélé} \end{cases}$$

Notre Statistique de test suit ici une loi du Chi-deux :

$$H = (\hat{b}_W - b^*)' [V(\hat{b}_W) - V(b^*)] (\hat{b}_W - b^*) \rightsquigarrow \chi^2(p-1)$$

Avec  $b^*$  l'estimateur des GLS.

En appliquant ce test à nos données nous obtenons le résultat suivant :

Table 14: Test H de Hausman

Stat.H	P.Value
385	0

Nous remarquons que pour notre test, nous rejettons l'hypothèse nulle  $H_0$ . Ainsi, nous concluons que nos effets individuels sont corrélés aux variables explicatives.

Avec ces 2 tests, nous mettons en évidence que les hypothèses liées aux estimateurs MCO, GLS ne sont pas respectées, ces estimateurs ne sont plus convergents, ainsi nous devons estimer différemment notre modèle. Cela montre qu'il manque des variables explicatives dans notre modèle (nous n'avons rien introduit d'autre pour le moment) mais que nous pouvons tout de même l'estimer d'où l'intérêt du panel dans ce cas.

Pour supprimer cette corrélation des effets individuels aux explicatives, la solution est d'utiliser l'estimateur *within* qui va supprimer les effets. En effet, il va retirer aux observations d'un département la moyenne de sa variable sur la période. Cela a pour effet de supprimer les effets car fixes au cours du temps mais aussi la constante de notre modèle. ( $\alpha_i - \frac{9 \times \alpha_i}{9} = 0$ )

Notre modèle (1) sera donc changé pour un modèle dans lequel pour chaque observation, la valeur des variables seront leur écart à la moyenne sur la période.

**8.3.2.1.2 Estimation du modèle** Avant d'estimer notre modèle, nous voulons alerter sur un point d'attention. Dans un modèle classique FEM il existe plusieurs hypothèses à respecter pour l'utiliser ou en tirer certaines inférences. Cependant, nous devons vérifier l'une d'entre elles qui est primordiale : l'indépendance des termes d'erreur. ( $Cov(\epsilon_{i,t}, \epsilon_{i,j}) = 0$  pour  $i \neq j$ )

Cela est dû au fait que dans le cadre des départements ici, une perturbation dans un département peut influencer les autres, il peut avoir des effets de voisinage. Cela va biaiser nos écart-types et donc nos inférences sur les coefficients. Par exemple, si les salaires sont hauts dans un département, cela peut influencer les habitants voisins à se déplacer, les politiques de certains départements peuvent affecter les autres, ... L'objectif ici est de vérifier ceci via le test de Pesaran.

En effectuant ce test, nous obtenons une p-value égale à 0.65. Ainsi, nos effets temporels et individuels ont l'air de bien capter toute les perturbations. Cependant, en effectuant un Test de Wooldridge sur panel, nous détectons de l'autocorrélation.

Les effets fixes annuels éliminent bien la composante macro purement nationale, mais ils ne suffisent pas à garantir l'indépendance des erreurs dans le temps ni entre départements. Afin de conserver des tests de significativité fiables, nous utiliserons les erreurs-types Driscoll-Kraay (De Hoyos and Sarafidis (2006)), qui restent valides même en présence d'autocorrélation et de dépendance croisée résiduelle (Hoechle (2007)). En faisant cela, nous garantissons la robustesse de nos résultats puisque le test de Pesaran est faiblement puissant lorsque T est faible comme dans notre cas. (Prudence) Concernant les tests de stationnarité (IPS, WU,...) notre panel a trop peu de périodes pour que les tests aient une puissance assez élevée. Nous estimons donc nos modèles aussi en différence première pour vérifier la robustesse de nos résultats. De plus, avec les séries temporelles nous avons déterminé qu'il y avait un équilibre de long terme et de la cointégration.

En estimant notre modèle simple FEM (1), nous obtenons les résultats suivants :

Coefficient	Estimate	Std. Error	p-value	Modèle
log_mutations	-0.0251437	0.0126071	0.0464766	FE simple
log_mutations	0.0041954	0.0046018	0.3622786	FE diff

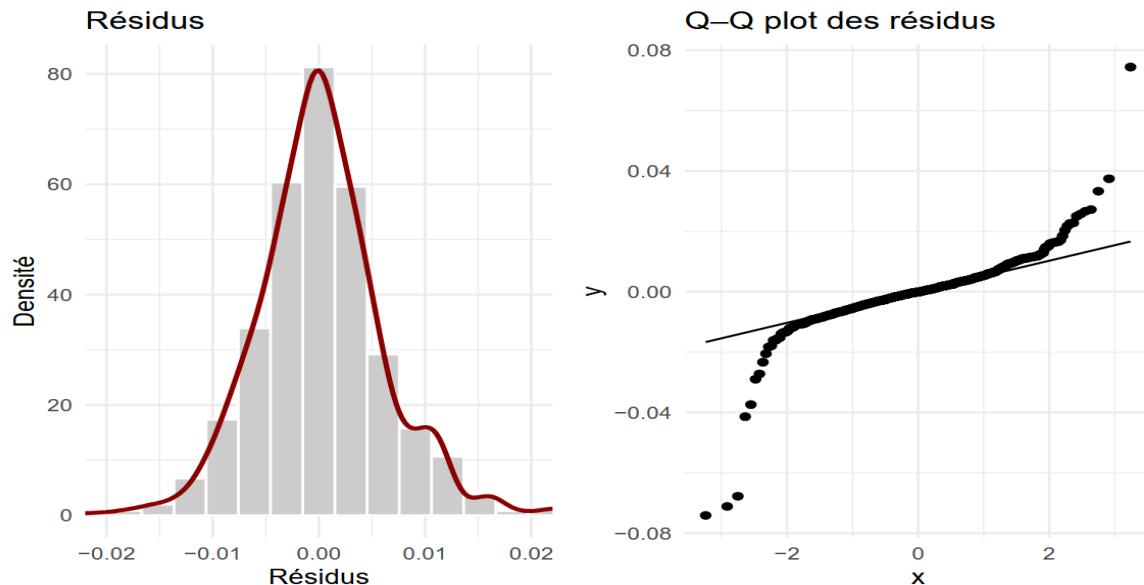
Nous remarquons que notre coefficient  $\beta$  est négatif et significatif (-0.025). IC 95% = [ -0,038 ; -0,012 ].

Ainsi, une hausse de 10% du nombre de transactions au sein d'un département baisserait de 0.25% le salaire net horaire moyen (énorme).

Nous sommes assez surpris d'avoir un coefficient négatif mais cela peut donc nous emmener à première vue sur l'hypothèse de la pression des offres d'emploi, l'afflux de main-d'œuvre. En effet, un marché immobilier très actif peut signaler une arrivée nette de nouveaux habitants ; si l'offre de travail augmente plus vite que la demande locale, la pression salariale baisse. Nous pourrions aussi observer un effet de composition, les secteurs à bas salaires (commerce, tourisme) connaissent un turn-over résidentiel plus élevé ; l'élasticité négative pourrait réfleter un changement de structure plutôt qu'un effet causal.

Il est encore trop tôt pour s'avancer autant. Retenons tout de même cette première valeur significative. De plus, en refaisant ce modèle sur le panel en différence première avec donc une période en moins ( $T=8$ ), nous obtenons un coefficient similaire, signe de la robustesse de celui-ci. L'effet négatif n'est donc pas lié à la non-stationnarité de nos séries.

L'analyse des résidus nous montre ceci :



Stat.test	p.value	Test
283.85	0	Breusch-Pagan
0.78	0	Shapiro-Wilk normality

Plusieurs analyses peuvent être faites sur notre modèle préliminaire. Ici, les résidus de notre modèle ne suivent pas une loi normale comme nous pouvions nous y attendre. Cependant, ils sont centrés sur 0 ce qui est une bonne nouvelle concernant le biais. En moyenne notre modèle est bon.

Le test de Brausch-Pagan nous confirme la présence d'hétéroscédasticité dans les erreurs. Cela nous amène à nous poser quelques questions sur l'estimation du coefficient. L'une des méthodes pour essayer de régler ce problème est de rajouter des variables dans notre modèle pour éviter les erreurs et obtenir une meilleure fiabilité des test de significativité.

**8.3.2.2 Modèle enrichi** La section précédente a établi que, toutes choses égales par ailleurs, une hausse de 10% du volume de transactions résidentielles est associée à une baisse d'environ 0,25% du salaire net horaire moyen. Ce résultat est stable d'ailleurs, quel que soit le traitement de stationnarité (niveaux ou différences premières). Cette corrélation brute constitue un point de départ, un comparatif, mais elle agrège potentiellement plusieurs canaux : conjoncture du marché du travail, structure sectorielle, composition démographique, ...

L'objectif de cette partie sera alors de purger notre variable d'intérêt : les mutations. Nous devons éviter les biais d'estimation et donc les régressions fallacieuses en contrôlant notre variable par d'autres qui pourraient potentiellement la composer. De plus nous allons essayer de chercher à savoir quelles sont les autres variables pouvant influencer les salaires ainsi que leur impact. Nous voulons tirer ici l'essence même de l'impact des mutations sur les salaires.

Pour ce faire, nous allons créer de nouveaux modèles en ajoutant peu à peu de nouvelles variables de contrôle et vérifier comment les coefficients bougent. Cela aura comme intérêt de nous renseigner sur la robustesse de nos résultats préliminaires ainsi que sur l'impact des différentes autres variables, quelles sont celles qui dominent ? La corrélation entre nos deux variables reflète-t-elle la structure sectorielle et démographique (territoires agricoles, vieillissants) plutôt qu'un mécanisme mobilité-salaire ?

Nos hypothèses sont qu'il y a forcément des effets de composition et que l'ajout de ces variables

de contrôle baissera partiellement notre coefficient mais pas au point de le ramener à 0. Si c'est le cas, alors cela montrera qu'il existe bien un impact propre aux mutations de logement.

Choix des variables de contrôle :

Dans l'ensemble de la section Panel, nous utiliserons les variables de contrôle suivantes:

Structure du marché du travail :

- Le taux de chômage, la conjoncture du marché du travail est importante à prendre en compte, le chômage peut varier énormément d'un département à l'autre ayant donc des conséquences sur son attractivité et donc la mobilité de la population. Lorsque le chômage baisse, une tension salariale s'intensifie ce qui facilite les mutations, il faut retirer cet effet des mutations. Dans notre panel cette variable est en point de pourcentage ( $chomage \in [0 : 100]$ ).
- Le nombre total d'emplois salariés, il mesure la taille du bassin d'emploi et donc cela permettra de neutraliser les effets sur les départements ayant des volumes important de mutations et salaires. L'objectif est d'éviter que le modèle nous indique qu'un département par défaut grand aura des salaires élevés. Cette variable est passée en logarithme pour les raisons évoquées dans la partie de présentation du panel.

Démographie :

- Part de la population jeune, cette variable renseigne sur la part des personnes agées de 25 à 39 ans. C'est cette tranche d'âge qui va être amenée à se déplacer, déménager le plus souvent. Elle peut donc avoir un effet sur les salaires et les mutations qu'il faut prendre en compte. Dans notre panel cette variable est en point de pourcentage ( $part25a39 \in [0 : 100]$ ).
- Part de la population viole, comme pour la variable précédente, celle-ci renseigne sur la part de la population agée de plus de 65 ans. Cet âge correspond généralement au passage à la retraite et peut influencer nos 2 variables assez logiquement puisque les salaires des retraités baissent, de plus ceux-ci peuvent vendre leur maison et changer de logement, etc... Dans notre panel cette variable est en point de pourcentage ( $part65 \in [0 : 100]$ ).
- Nous avons aussi ajouté comme mentionné plus haut, une variable dichotomique *ruralité* qui vaut "1" si le département est dit rural (moins de 100 hab/km<sup>2</sup>), 0 sinon.

Composition sectorielle :

- Part d'emplois salariés dans le secteur Agricole, cette variable peut être intéressante à ajouter puisque ces métiers ont tendance à être des pratiqués en ruralité, et à salaire faible cela peut donc jouer dans le modèle. Nous allons pouvoir neutraliser les écarts salariaux en fonction de la structure sectorielle du département. Dans notre panel cette variable est en point de pourcentage ( $partAgricole \in [0 : 100]$ ).
- Part d'emplois salariés dans le secteur Industriel, même raisonnement bien que nous nous attendions ici à ce que celui-ci ait un effet positif sur les salaires.
- Part d'emplois salariés dans le secteur Tertiaire, même spécialisation que précédemment bien qu'il faut noter que celle-ci concerne seulement le tertiaire marchand. Bien que nous ne prenons pas le non-marchand, cela ne pose pas de problème puisque par la suite pour éviter tout problème de multi-colinéarité nous retirerons cette variable qui sera donc prise en référence dans le modèle en quelque sorte.
- Part d'emploi dans le secteur de Construction, ici même raisonnement que pour le secteur agricole.

Immobilier :

- La surface moyenne des logements, ici cela rejoint un peu le contraste entre ruralité et non ruralité dans le sens où nous nous attendons à ce que dans les territoires ruraux, il y ait une proportion de maison plus grande que dans les départements non-ruraux. Nous le montrerons par la suite.
- Le prix au m<sup>2</sup> des logements, ceci va agir un peu comme un indice des prix et va nous permettre de purger cet aspect là. Il peut être à la fois plus difficile d'acquérir des logements s'ils ont un prix élevé mais cela peut aussi refléter une certaine attractivité de la région et donc une offre supérieure à la demande et des prix élevés et donc ainsi des salaires aussi plus élevés.
- On ajoute aussi la proportion de maison dans les mutations.

Ces variables seront donc les seules utilisées dans nos modèles.

Nous aurions pu incorporer bien d'autres variables de contrôles qui auraient un sens économique, nous pouvons notamment parler de l'inflation pour les salaires avec l'IPC mais dans le cadre du modèle within il n'y aurait pas de variation de par le caractère commun de l'inflation au niveau national. Il est déjà pris en compte dans nos effets temporels donc ça n'aurait pas de sens de le rajouter, de même pour les taux d'emprunt et autres. Un autre point pertinent aurait été celui de la composition par catégorie socioprofessionnelle pour expliquer les salaires puisque les cadres ont naturellement des salaires plus élevés que les ouvriers et employés mais nous n'avons pu les considérer puisque "les Bases Tous salariés" ne sont pas toutes publiques pour les années que nous considérons.

Nous n'utilisons pas les autres variables de la base DVF puisque celles-ci sont logiquement calculées sur l'échantillon de logement qui d'est vu changer de propriétaire donc ne représentant pas les véritables caractéristiques du parc immobilier de chaque département. De plus ces 2 variables sont plus ou moins liées et peuvent interagir avec des effects conjoncturels et la valeur moyenne des surfaces ou nombre de pièces sont souvent une conséquence et non pas une cause.

#### 8.3.2.2.1 Résultats des modèles

Le modèle retenu est le suivant :

$$s_{i,t} = X_{i,t}\hat{\beta} + \alpha_i + \delta_t + \epsilon_{i,t} \quad (2)$$

Avec :

- $X_{i,t}$  le vecteur de dimension (1,7) pour un département  $i$ . Ce vecteur est composé de l'ensemble de nos variables explicatives retenues, (*log\_mutations, Chomage, part\_25\_39, part\_65\_plus, p\_industrie, p\_agricole, p\_construction*).
- $\hat{\beta}$  le vecteur estimé de nos paramètres, coefficients associés aux variables explicatives. Il est de dimension (7,1)

Table 17: Coefficients et p-values des modèles en niveau

Variable	FEM 1	FEM 2	FEM 3	FEM 4
log_mutations	-0.014 (0.189)	-0.015 (0.147)	-0.016 (0.095)	-0.019 (0.048)
Chomage	0.004 (0.059)	0.004 (0.059)	0.003 (0.036)	0.001 (0.543)
log_tout_emploi	0.139 (0.000)	0.139 (0.000)	0.147 (0.000)	NA
part_25_39	NA	-0.063 (0.892)	-0.133 (0.777)	-0.266 (0.584)
part_65_plus	NA	0.003 (0.988)	-0.088 (0.673)	-0.464 (0.028)
log_pop	NA	-0.000 (0.973)	-0.000 (0.988)	NA
p_Industrie	NA	NA	0.002 (0.132)	0.003 (0.002)
p_Construction	NA	NA	-0.007 (0.066)	-0.005 (0.158)
p_Agriculture	NA	NA	-0.009 (0.046)	-0.010 (0.022)
p_Tertiaire	NA	NA	-0.001 (0.628)	NA

Dans le modèle retenu, nous avons retiré certaines variables à cause de la colinéarité après un VIF.

Le modèle enrichi met en évidence une élasticité négative (-0.019) mais de faible ampleur du salaire net horaire moyen au volume de transactions. Cette relation persiste après neutralisation de la structure du marché de l'emploi, de la structure d'âge et de la composition sectorielle, suggérant que les mutations auraient un impact négatif sur les salaires nets horaires moyens à court terme. L'effet apparaît plus marqué dans les territoires vieillissant ou à dominante agricole, tandis que la densité industrielle compense partiellement cette tendance.

Nous pourrions ici penser que notre hypothèse initiale sur l'afflux d'offre de travail se comfirme, que l'afflux pourrait être lié aussi à la saisonnalité, etc... puisque l'élasticité reste négative et a réduit d'une faible ampleur montrant une certaine robustesse à première vue mais il faut faire preuve de prudence, nos résultats sonnent assez fallacieux, le chômage ici n'a pas d'impact sur les salaires cela est assez étrange, de plus, les estimations de ces modèles sur les variables en différence première, nous donnent ceci :

Table 18: Coefficients et p-values des modèles en différence première

Variable	FEM 1	FEM 2	FEM 3	FEM 4
log_mutations	0.005 (0.265)	0.006 (0.183)	0.006 (0.243)	0.006 (0.241)
Chomage	0.003 (0.055)	0.003 (0.026)	0.003 (0.027)	0.003 (0.035)
log_tout_emploi	-0.019 (0.561)	-0.018 (0.563)	0.049 (0.500)	NA
part_25_39	NA	-0.577 (0.062)	-0.581 (0.060)	-0.572 (0.065)
part_65_plus	NA	0.433 (0.508)	0.467 (0.477)	0.456 (0.477)
log_pop	NA	0.001 (0.268)	0.001 (0.272)	NA
p_Industrie	NA	NA	0.001 (0.875)	0.002 (0.525)
p_Construction	NA	NA	0.001 (0.806)	0.003 (0.241)
p_Agriculture	NA	NA	-0.001 (0.936)	0.000 (0.962)
p_Tertiaire	NA	NA	-0.002 (0.549)	NA

Nous remarquons ici que tout s'écroule, nos mutations n'ont plus d'effet significatif et certaines variables ont changé de signe. Ainsi, nous pouvons penser que l'élasticité négative que nous trouvions captée en niveau est en réalité d'origine structurelle. La différenciation supprime les tendances longues communes ; si la relation était réellement causale à court terme, on la retrouverait dans le modèle en différence. Le fait qu'elle devienne nulle et même légèrement positive (non significative) indique qu'elle provenait de caractéristiques permanentes (densité, spécialisation, urbanité) déjà contrôlées par les effets fixes mais encore corrélées à la tendance propre des séries.

Nous pouvons même nous questionner sur le risque de regression fallacieuse ici. L'exercice en différences confirme que la baisse salariale observée dans les départements très actifs sur le marché immobilier ne se matérialise pas instantanément : il s'agit plus d'un écart de niveau de long terme que d'un ajustement immédiat.

Le salaire devient significatif dans ce modèle, ce qui semble bien plus cohérent tout de même avec la théorie économique. Il est aussi normal d'observer un effet négatif sur la tranche d'âge 25-39 ans puisqu'ils sont en début de carrière et vont en général avoir des salaires plus bas.

Pour conclure cette partie de l'impact des mutations sur les salaires, nous pouvons dire que lorsque nous passons en différences premières, l'élasticité négative des salaires au volume de mutations

disparaît, tandis que la variation du chômage émerge comme déterminant principal de la dynamique salariale. Ces résultats suggèrent que l'association négative repérée en niveau relève davantage de différences structurelles, de la densité, spécialisation, démographie que d'un mécanisme de court terme. Il faut faire attention avec ces résultats puisqu'ils sont calculés seulement avec une période T=9 ou 8 (diff) ce qui augmente le bruit.

Les analyses concernant la ruralité sont vaines de par le fait que nous disposons de trop peu de données et les périodes étant courtes, nous ne pouvons sortir de résultat concret.

Toutes les analyses sur les résidus de ces modèles sont en annexe. Pour obtenir des résultats plus robustes il faudrait plutôt se diriger vers un modèle ECM. Nos variables suivent certainement des tendances communes comme vu dans la partie série temporelle d'où le besoin d'agir sur celles-ci avant de faire nos modèles pour éviter les risques de régressions fallacieuses comme nous avons pu l'avoir précédemment. Nous trouvons qu'il existe de la cointégration ainsi nous pouvons construire notre modèle.

Notre Modèle de Panel ECM sera le suivant :

$$\Delta s_{i,t} = \gamma(s_{i,t} - \beta' X_{i,t-1}) + \phi' \Delta X_{i,t} + \alpha_i + \delta_t + \epsilon_{i,t} \quad (3)$$

Avec :

- $\delta X_{i,t}$  les variations de nos variables explicatives, cours terme.
- $\gamma$  La vitesse de convergence associée à l'équilibre de long terme.

Les interprétations sont les mêmes que dans la partie série temporelle donc nous ne reviendrons pas dessus ici.

Table 19: Estimation Panel ECM  $R^2 = 0.4538$

Variables	Coefficient	P_value
equilibre_LT	-0.8833	0.0000
d_log_mutations	-0.0012	0.6826
d_Chomage	0.0031	0.0182
d_log_prixm2	0.0401	0.0008

Variables	Coefficient	P_value
d_part_25_39	-0.8715	0.0246
d_part_65_plus	-0.2482	0.6613
d_p_Agriculture	-0.0058	0.5077
d_p_Industrie	0.0034	0.0683
d_p_Construction	-0.0090	0.0085

Relation de long terme :

Le coefficient lié à l'équilibre de long terme est de -0,8833. Ainsi, si le salaire net horaire moyen observé s'écarte de son niveau d'équilibre fondé sur les mutations immobilières et nos autres variables, alors 88 % de cet écart sera corrigé l'année suivante. Économiquement, nous interprétons cela ainsi : les salaires départementaux évoluent fortement vers le prix d'équilibre déterminé par le marché immobilier et la composition sectorielle. Un dysfonctionnement (salaire trop élevé ou trop bas) est presque intégralement dissipé en un an, ce qui traduit une forte flexibilité salariale de court terme face aux tensions du marché résidentiel.

Relations de court terme :

- Chômage : Une hausse de 1 point de pourcentage du taux de chômage d'une année à l'autre est associée à une progression de +0,31% du salaire moyen l'année suivante. Bien que contre-intuitif à première vue, cet effet peut refléter un effet compositionnel : en phase de hausse du chômage, ce sont souvent les postes les moins qualifiés qui se réduisent, tirant la moyenne salariale vers le haut.
- Prix au m<sup>2</sup> : Conformément aux attentes, une augmentation de 1% des prix immobiliers se traduit rapidement par +0,04% de hausse salariale. Cela peut signaler un effet de coût de la vie : face à la hausse des loyers et des emprunts, les négociations salariales intègrent une prime protectrice.
- Part des 25-39 ans : L'accroissement de 1 point de pourcentage de la part des 25-39 ans correspond à une baisse de 0,87% de la croissance salariale. Ce résultat suggère qu'un flux accru de jeunes actifs peut exercer une pression à la baisse sur les salaires moyens, via un effet d'offre de travail plus abondante dans cette tranche d'âge.

- Part de la construction : Une hausse d'un point de la part des emplois en construction pèse faiblement mais significativement sur la croissance salariale, conformément à la plus faible rémunération dans ce secteur.

Les autres variations sectorielles (agriculture, industrie, senior 65+, etc.) n'apparaissent pas significatives à court terme, ce qui indique que leur impact annualisé reste marginal comparé aux dynamiques de chômage, de prix et de structure démographique. Nous n'observons pas d'effet de court terme des log\_mutations sur les salaires.

A la suite des ces analyses, nous rappelons que ces résultats sont à reconsidérer pour de nombreuses raisons. Nos estimations peuvent être biaisées à cause du nombre de périodes restreintes ( $T=8$ ). Nous avons vérifié la cointégration via Pedroni en sortant un vecteur de cointégration pour chaque département or il est pratiquement sûr qu'en réalité celui-ci diffère un peu d'un département à l'autre (Nous ne pouvons le tester sur d'autres échantillons par manque d'observations). Nous avons remarqué aussi qu'il reste un peu d'autocorrélation spatiale à la suite d'un test de Pesaran. Ici une approche plus locale serait préférable, le manque de granularité se fait ressentir dans les résultats.

### 8.3.3 Sous annexe Panel

Retrouvez ici tous les résultats cités précédemment et non intégrés dans les parties.

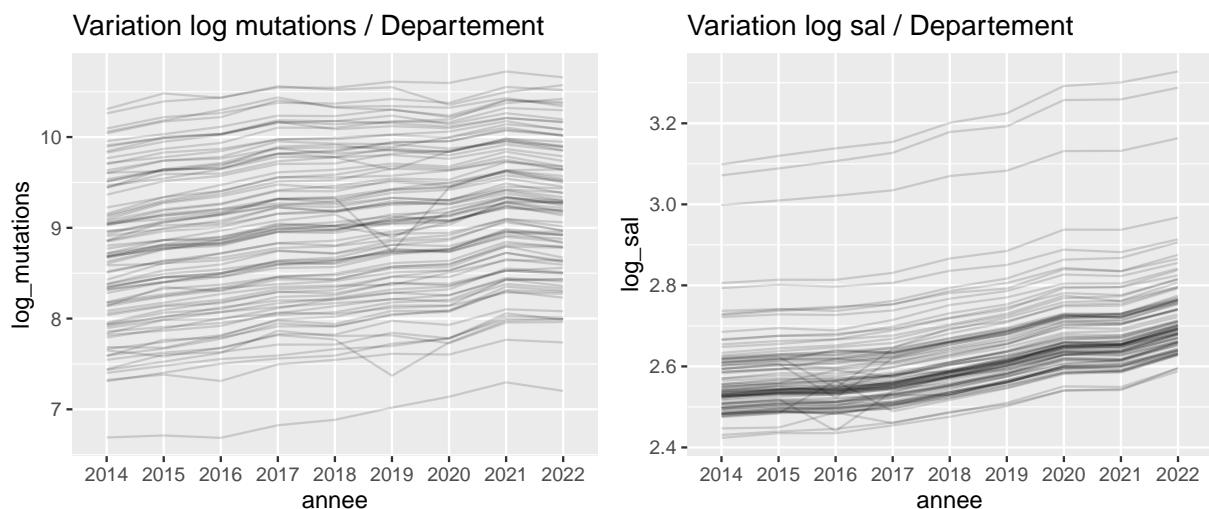


Table 20: VIF Panel

	x
log_sal	4.775920
p_Industrie	1.748673
p_Agriculture	1.476627
p_Construction	1.414970
Chomage	1.650977
part_25_39	2.683615
log_prixm2	4.926212

Table 21: Correlations salaire mutations au cours des années

annee	corr
2014	0.6480573
2015	0.6417096
2016	0.6157390
2017	0.6310054
2018	0.6317443
2019	0.6260169
2020	0.6028587
2021	0.6033588
2022	0.6091275

RURAL VS NON-RURAL :

Modele enrichi Salaires ~ Mutations

Les graphiques et tests :

log_mutations -0.004 (0.009)	
Num.Obs.	504
R2	0.000
AIC	-3371.0
BIC	-3362.6
RMSE	0.01
Std.Errors	Custom

Figure 44: Analyse modele rural

log_mutations -0.032 (0.020)	
Num.Obs.	333
R2	0.062
AIC	-2276.6
BIC	-2269.0
RMSE	0.01
Std.Errors	Custom

Figure 45: Analyse modele non rural

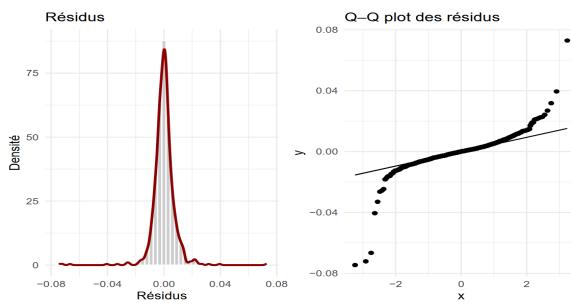


Figure 46: Analyse résidus du modèle enrichi

Table 22: Variables en niveau

Stat.test	p.value	Test
386.53	0	Breusch-Pagan
0.77	0	Shapiro-Wilk normality

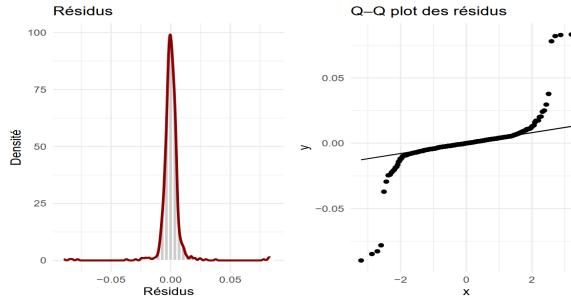


Figure 47: Analyse résidus du modèle enrichi diff

Table 23: Variables en différence première

Stat.test	p.value	Test
71.00	0	Breusch-Pagan
0.53	0	Shapiro-Wilk normality

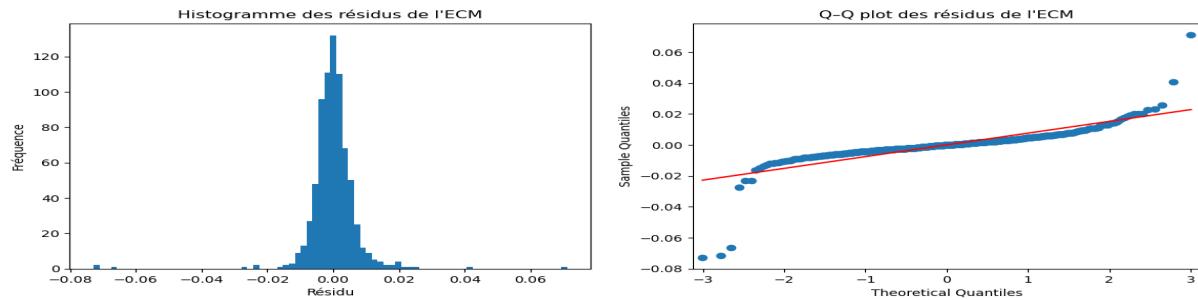


Table 24: Test Pedroni Cointegration Panel

	empirical	standardized
nipanel	6.620158	-15.598354
rhopanel	-105.240056	3.728268

	empirical	standardized
tpanelnonpar	-38.512807	-15.621773
tpanelpar	-5305.192173	-6355.958658
rhogroup	-113.101739	6.664652
tgroupnonpar	-43.481086	-21.187754
tgrouppar	-41.931063	-19.107140

Parameter Estimates						
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
equilibre_LT	-0.7633	0.1194	-6.3917	0.0000	-0.9978	-0.5288
d_log_mut	-0.0110	0.0037	-2.9628	0.0032	-0.0182	-0.0037

Figure 48: Estimation du modèle ECM simple

Modèle enrichi Mutations ~ Salaires

Les graphiques et tests :

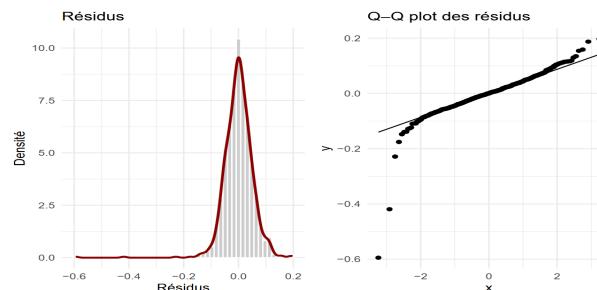


Figure 49: Analyse résidus modèle enrichi

Table 25: Variables en niveau

Stat.test	p.value	Test
41.431	0	Breusch-Pagan
0.870	0	Shapiro-Wilk normality

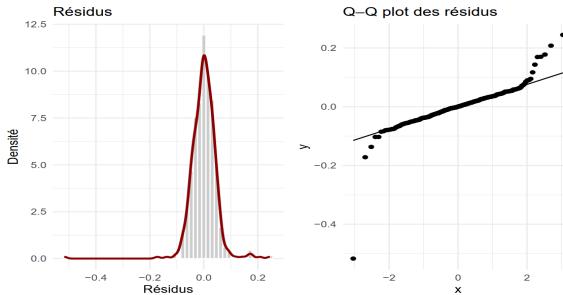


Figure 50: Analyse résidus modèle enrichi diff

Table 26: Variables en différence première

Stat.test	p.value	Test
72.724	0	Breusch-Pagan
0.810	0	Shapiro-Wilk normality

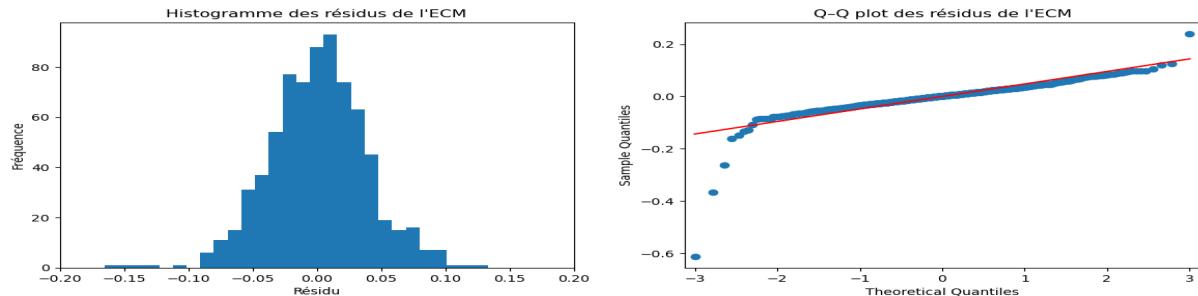


Table 27: Test Cointegration Panel ECM mut-sal

	empirical	standardized
nipanel	1.641074	-10.925464
tpanelnonpar	-35.668767	-14.946034
tpanelpar	-4360.954809	-4420.254429
rhogroup	-93.898108	4.302249
tgroupnonpar	-37.438752	-17.532573
tgrouppar	-35.960682	-15.652387

Parameter Estimates						
Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI	
equilibre_LT	0.1229	-4.3460	0.0000	-0.7752	-0.2927	
d_log_sal	0.1554	-1.0000	0.3177	-0.4606	0.1498	

Figure 51: Estimation du modèle ECM simple