INSTITUT DES ACTUAIRES

Lyon 1

ISFA GRANDE ÉCOLE D'ACTUARIAT ET DE GESTION DES RISQUES
INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Souhail El Ansari

Titre : Analyse du risque inondation aux États-Unis : Étude du *National Flood Insurance Program* et modélisation de l'autocorrélation spatiale des taux de destruction.

Confidentialité : NON ☐ (Durée : ☐ 1 an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires — Signature

.....

.....

.....

Membres présents du jury de l'ISFA

.....

.....

.....

Entreprise :
Nom : Seabird Conseil
Signature :

Directeur de mémoire en entreprise :
Nom : Pierre Thérond
Signature :

Invité :
Nom :
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

S. El

Résumé

Les montants engendrés par les catastrophes naturelles sont en constante augmentation et sont amenés à poursuivre leur croissance en raison des effets liés au changement climatique. Le risque inondation pose de nombreux défis aux acteurs publics et privés pour concevoir un système viable financièrement. L'ampleur des pertes possibles, la volatilité du risque, la difficulté à créer un pool suffisamment diversifié sont autant de problèmes posés par ce risque.

Aux États-Unis, l'absence de couverture assurentielle, à la suite du retrait des assureurs privés après les grandes inondations du Mississippi en 1927, a entraîné le Congrès à adopter en 1968 le National Flood Insurance Program, à savoir un programme d'assurance inondation où le risque est porté par le gouvernement fédéral. Ce système fait face à de nombreuses critiques en raison d'une dette difficile à résorber ou d'une demande faible des propriétaires pour ce type de couverture. Dans une démarche de transparence, et afin de mieux comprendre ce risque, la *Federal Emergency Management Agency*, l'organisme gouvernemental américain qui assure la protection face aux grandes catastrophes naturelles, a publié en juin 2019 une base de données des sinistres sur un historique de 40 ans couverts par le *National Flood Insurance Program*.

Nous ferons une étude des outils de gestion du risque inondation au niveau mondial afin de comprendre les enjeux et les moyens utilisés par les États et assureurs privés pour gérer le risque financier lié aux inondations. Nous ferons un gros plan sur ce risque aux États-Unis où nous étudierons grâce à la base du NFIP la sévérité du risque aux États-Unis. Dans un premier temps nous tenterons de modéliser le montant des dommages grâce à des régressions Gamma et Lognormale. La prime payée par les assurés étant un pourcentage de la valeur assurée, nous étudierons, dans un second temps, le taux de destruction grâce à un modèle de régression Bêta proposée par Ferrari et Cribari-Neto (2004).

Dans un dernier temps, nous nous intéresserons à l'implémentation des modèles issus de l'économétrie spatiale, qui permettent de tenir compte de l'autocorrélation spatiale des observations. Nous verrons comment détecter ce phénomène grâce à l'indice de Moran. Après avoir explicité les manières de définir les relations de voisinage d'un ensemble de zones, nous appliquerons un modèle spatial autorégressif sur le taux de destruction à l'échelle des comtés et verrons ses améliorations par rapport à une régression linéaire classique et quelles en sont les limites.

Mots clés : inondation, GLM, régression spatiale, taux de destruction, NFIP, régression Bêta.

Abstract

Cost of natural disasters are rising and are likely to keep on rising because of climate change. Flood risk is a significant challenge for public and private stakeholders when trying to manage its financial impact. Magnitude and volatility of the losses or the difficulty of creating a sufficient large pool of diversified risk pose challenges to insuring flood risk.

In the United States, lack of flood risk insurance in the wake of 1927 Mississippi floods led the Congress to create the National Flood Insurance Program to provide flood insurance covered by the Federal Government. This system is under fierce criticism because of a high debt and a low take-up rate. In order to promote transparency and to have a better understanding of flood risk, the Federal Emergency Management Agency released in June 2019 released NFIP data that includes two million claims records dating back to 1978.

We will try to understand, throughout the world, the challenges to insuring flood risk and how States and private insurers deal with this risk. We will focus on the United States thanks to NFIP data to understand risk severity. First we will use Gamma and Lognormal models to explain the value of losses. As the premium is a percentage of the value insured, we will try to model the destruction rate with the Beta regression model proposed by Ferrari and Cribari-Neto (2004).

In a final part, we will understand how to implement models from spatial econometrics, to take into account spatial autocorrelation of the values. We will see how we can detect it thanks to Moran Index. After explaining how we can define neighborhood relations between areas, we will model rate destruction at county-level and how they can fit better than linear regression. Those spatial models have some limits that will be detailed.

Keywords : flood, GLM, spatial regression, destruction rate, NFIP, Beta regression.

Remerciements

Ce mémoire est l'aboutissement de mes études, je remercie toutes les personnes qui m'ont aidé à le réaliser.

Je tiens à remercier Pierre Thérond et Yahia Salhi pour leurs précieux conseils.

Je remercie l'ensemble de mes camarades de promotion pour leur soutien.

Table des matières

Introduction

Aux États-Unis, l'année 2019 a été marquée par d'importantes inondations qui ont touché les États du Midwest et le sud des États-Unis. Près de 14 millions de personnes ont été affectées par celles-ci. Le montant des dommages est estimé à 20.3 milliards $ (NOAA).

L'inondation est l'une des catastrophes naturelles les plus destructrices dans le monde. Elles affectent chaque année près de 250 millions de personnes et ses coûts financiers sont évalués à 40 milliards $(OCDE, UNISDR).

Dans de nombreux pays, une partie importante de la population vit en zone inondable : les deux-tiers de la population néerlandaise vivent dans une zone à risque. Au Japon 49 % de la population vit dans un ancien cours d'eau. Les deux tiers des déclarations de catastrophes naturelles aux États-Unis entre 1953 et 2010 sont dues aux inondations (OCDE, 2016).

A cela s'ajoute le changement climatique, qui pourrait être à l'origine d'une augmentation du risque d'inondation. Dans les zones terrestres, la fréquence des fortes précipitations s'intensifieront d'après certaines études (Keating et al, 2014), ce qui augmentera la fréquence des crues. Dans les zones côtières, l'augmentation du niveau des mers, de l'intensité des cyclones tropicaux et ouragans entraînent également une hausse du risque, à la fois en terme de fréquence et de sévérité.

Pour faire face aux coûts énormes engendrés par les inondations, de nombreux moyens ont été mis en œuvre pour gérer le risque financier, prévenir le risque et améliorer la résilience des populations touchées. Les mesures de prévention peuvent être la construction de digues, l'interdiction des permis de construire en zone inondable, des campagnes de sensibilisations ou l'élévation des maisons. Les gouvernements et les assureurs privés ont développé des outils de transfert du risque financier mais la question qui se pose est de savoir quel est le moyen le plus efficient pour protéger les populations.

Aux États-Unis, le programme national d'assurance inondation (*National Flood Insurance Program*, NFIP) a été créé en 1968, en raison du retrait des assureurs privés, avec deux objectifs : permettre aux propriétaires d'obtenir une assurance inondation soutenue par le gouvernement fédéral à travers un mecanisme de partenariat public-privé, et encourager les états à limiter la construction d'infrastructure et habitations en zone inondable. Cependant, la demande d'assurance inondation reste faible même dans les zones où elle est obligatoire. Le coût des primes ou la sous-estimation du risque peuvent expliquer cette faible demande (Kousky et Shabman, 2015).

La nature même du risque d'inondation fait qu'il y a de nombreux obstacles qui peuvent entraver son développement. Par exemple, les inondations sont localisées

dans des points précis, c'est-à-dire autour des fleuves et rivières et sur les zones côtières : il y a une corrélation spatiale de la sinistralité.

Le but de ce mémoire est, après avoir fait une revue des outils de gestion du risques au niveau mondial, de s'intéresser à la sévérité du risque inondation aux États-Unis et à l'autocorrélation spatiale de ce risque. Dans un premier temps nous tenterons de modéliser le montant des dommages en étudiant les résultats issus d'une régression Gamma et lognormale. La prime payée par les assurés étant un pourcentage de la valeur assurée, nous étudierons, dans un second temps, le taux de destruction grâce à un modèle de régression Bêta proposée par Ferrari et Cribari-Neto (2004). Dans un dernier temps, nous nous intéresserons à l'implémentation des modèles issus de l'économétrie spatiale, qui permettent de tenir compte de l'autocorrélation spatiale des observations. Nous utiliserons les méthodes présentées par l'Insee dans son *Manuel d'analyse spatiale* (2018) pour les appliquer à notre étude sur le taux de destruction à l'échelle des comtés.

1 Chapitre I : Présentation du risque inondation

1.1 Gestion du risque

1.1.1 Systèmes de couverture soutenus par l'État

En France, les assurances dommages aux biens et pertes d'exploitations sont dans l'obligation d'inclure une garantie contre les risques naturels d'après l'article L 125-1 : *Les contrats d'assurance, souscrits par toute personne physique ou morale autre que l'État et garantissant les dommages d'incendie ou tous autres dommages à des biens situés en France, ainsi que les dommages aux corps de véhicules terrestres à moteur, ouvrent droit à la garantie de l'assuré contre les effets des catastrophes naturelles, dont ceux des affaissements de terrain dus à des cavités souterraines et à des marnières sur les biens faisant l'objet de tels contrats En outre, si l'assuré est couvert contre les pertes d'exploitation, cette garantie est étendue aux effets des catastrophes naturelles, dans les conditions prévues au contrat correspondant.* La CCR récolte une prime additionnelle proportionnelle sur chaque contrat (12 %). Elle dispose de la garantie illimitée de l'État. Une intervention de réassureurs privés est possible sans toutefois remettre en cause le versement de la prime additionnelle à la CCR. La surprime ne dépend pas du niveau de risque de l'assuré. Cependant la franchise peut être modulable et inciter l'assuré à prendre des mesures de prévention. Cette franchise est déterminée par l'État et augmente en fonction du nombre de reconnaissance de l'état de catastrophe naturelle d'une zone. Cette reconnaissance est impérative pour obtenir une indemnisation de l'assureur.

Les catastrophes naturelles en Espagne sont couvertes par une entreprise publique le Consorcio de compensacion de Seguros. À la différence de la France, la couverture des risques naturels est intégrée dans les polices d'assurances aux biens, vie et accidents personnels. Une surprime est collectée par les assureurs privés qui la reversent au Consorcio. Il n'y a pas de franchise sur les dommages personnels. Dans ce modèle, l'intervention de l'État se fait directement au niveau de l'assurance et non à celui de la réassurance comme en France. Le consorcio peut transférer une partie de son risque au réassureur et bénéficie du soutien illimité de l'État.

En Nouvelle-Zélande les risques naturels sont couverts grâce aux polices d'assurance incendie souscrites auprès d'assureurs privés. Les primes additionnelles sont touchées par le National Disaster Fund, administré par un organe public ayant la garantie de l'État. La prime est proportionnelle à la valeur du bien assuré (15 cents pour chaque 100$ NZ assuré pour les immeubles résidentiels avec une limite de 207 $NZ par an). (Sénat, 2017). Contrairement à la France ou l'Espagne où l'assurance la garantie contre les risques naturels est illimitée, en Nouvelle-Zélande celle-ci est limitée à 100 000$ NZ pour les immeubles et 20 000$ NZ pour les meubles. La partie supérieure peut être couverte par des assureurs privés : le nombre élevé de catastrophes naturelles et la faible récolte de primes mettrait à mal l'équilibre financier du système si la garantie n'était pas limitée.

En Belgique, comme en Nouvelle-Zélande, la couverture des catastrophes naturelles est liée à une police d'assurance incendie. Il n'y a cependant pas d'assureur ni de réassureur public. Le bureau de tarification belge, composé par des assureurs et consommateurs, définit les plafonds tarifaires des primes couvrant les risques de catastrophes naturelles. En cas de sinistre, l'assureur prend en charge l'indemnisation jusqu'à un certain montant. La charge restante est répartie entre tous les assureurs présents en Belgique selon une répartition fixée par la Caisse de compensation des catastrophes naturelles. L'indemnisation des sinistres est, comme en Nouvelle-Zélande, plafonnée. Si l'état de calamité publique est déclarée, la Caisse des calamités prend en charge la somme restante. (Sénat, 2017)

1.1.2 Concertation entre assureurs privés et État

Au Royaume-Uni, la couverture des inondations est opérée par les assureurs privés, sous un régime négocié avec les institutions publiques. Les acteurs (gouvernement et assureurs), considérant que la forte exposition des propriétés aux inondations et le montant croissant des primes constituaient un frein pour les ménages pauvres, ont établit un nouveau système de réassurance Flood Re. Celle-ci a été institué par le *Water act 2014*. Cette réassurance est accessible aux assureurs pour limiter leur exposition au risque d'inondation. Les assureurs ont le choix de céder les polices qu'elles souhaitent à Flood Re. Les primes de réassurance couvrant le risque inondation sont indexés sur la valeur de la propriété, cela dans le but de maintenir des primes abordables pour les assurés à faibles revenus.

1.1.3 Assureurs privés sans obligations légales

Dans de nombreux pays, il n'existe pas de mécanisme d'assurance contre les risques naturels obligatoires ou liées de facto à une autre assurance. Prenons l'exemple de l'Italie où il n'existe aucun dispositif de réassurance publique couvrant les assureurs privés en cas de perte. L'assurance se fait sous la base du volontariat auprès d'assureurs privés. La prime est liée au risque, ce qui entraîne un coût élevé, voire rédhibitoire, pour les assurées dans les zones les plus exposées et une sélection adverse. Ce phénomène associé à une faible protection au sein des garanties (couverture des pertes directes uniquement) font que le taux de pénétration du marché de l'assurance inondation en Italie est particulièrement faible.

En Allemagne, les risques naturels sont couverts exclusivement par les assurances privées, sans appui de la puissance publique. L'assurance inondation est en option sur les contrats d'assurance habitation classiques. Le coût élevé de l'option freine les assurés : en 2013, seuls 35 % des propriétaires avaient souscrit à une assurance inondation. Les inondations en 2002 à Dresde ont conduit les assureurs allemands à augmenter les primes de 60 % dans les zones à risque (OCDE, 2016).

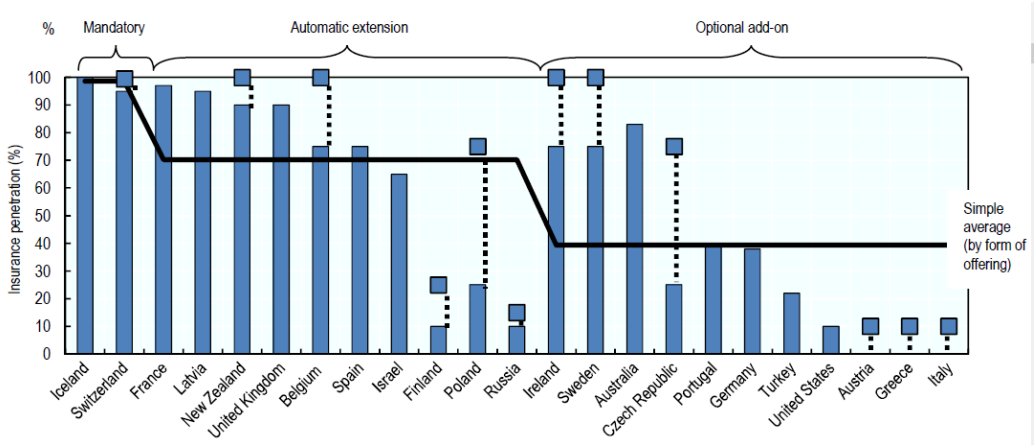


FIGURE 1: Taux de pénétration de l'assurance inondation, OCDE, 2016

1.1.4 Microassurance

Dans de nombreux pays en développement, en raison d'une faible offre d'assurance et d'un prix rédhibitoire pour une grande partie de la population, le taux de pénétration est très faible. Pour palier ce phénomène, des projets de micro-assurance ont vu le jour afin de proposer des couvertures à des prix abordables. La couverture du risque inondation dans le domaine de la microassurance se fait généralement au travers d'une assurance paramétrique. Au Pérou par exemple, l'assureur La Positiva propose une assurance paramétrique indexée sur la température de la mer. Le phénomène El Niño modifie la température de la surface de la mer et bouleverse les conditions météorologiques. Il s'en suit des tempêtes et de fortes pluies. La couverture paramétrique permet d'indemniser les assurés avant l'arrivée des pluies et de financer des mesures de prévention.

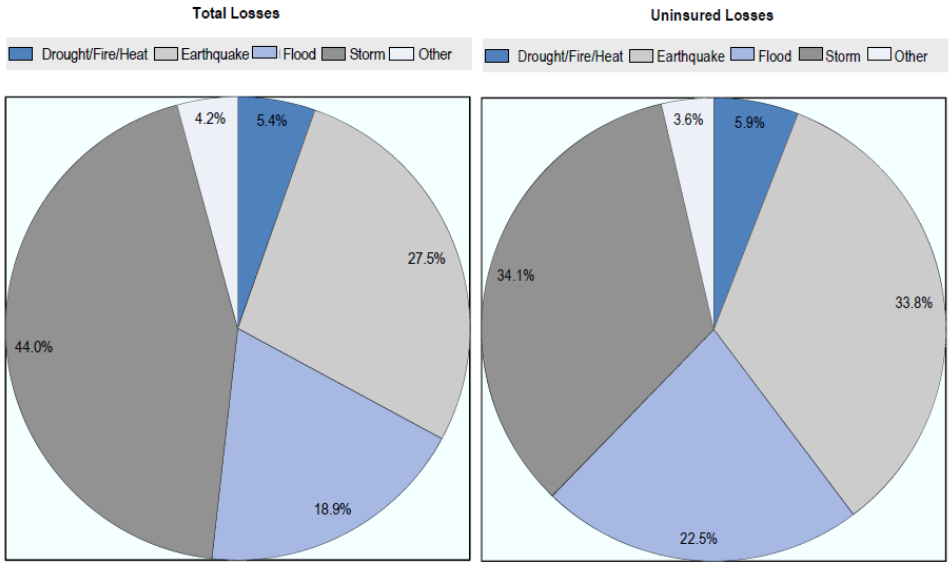
1.1.5 Aide publique

L'impact financier énorme des inondations peut ne pas être absorber simplement par une couverture d'assurance. Dans de nombreux pays, l'État intervient pour financer les pertes non assurables. Au Canada, la couverture du risque inondation est récente et n'est pas disponible pour toutes les propriétés (OCDE, 2016). L'État intervient à travers un fond d'urgence pour venir en aide aux propriétaires sinistrés.

1.2 Défis posés par l'assurance inondation

Malgré les initiatives privées et publiques proposant une couverture d'assurance, une part non négligeable de propriétaire habitant dans des zones à risque ne se couvre pas contre le risque inondation. Ce graphique nous montre que la part des

inondations dans les pertes totales liées aux catastrophes naturelles est de 18.9 % entre 2005 et 2015 mais 22.5 % des pertes non assurées (Figure 2).



Notes: It should be noted that the figures take account of losses insured by the National Flood Insurance Program (i.e. the share of insured losses includes losses insured by the NFIP).

Source: OECD calculations based on insured losses and total damages reported for natural disasters (floods, storms, earthquake, droughts/fires/heat waves and other natural disasters) in Swiss Re sigma annual reports on natural and man-made catastrophes (2005-2015).

FIGURE 2: Part des pertes non assurées, OCDE, 2016

Comme nous l'avons vu dans le graphique précédant le taux de pénétration varie selon les pays. On remarque que dans les pays où la couverture est facultative, le taux de pénétration est faible. En Irlande et en Suède, la couverture inondation est obligatoire si l'on veut obtenir un prêt immobilier. Cela explique les forts taux de souscription. Cependant, aux États-Unis, malgré le fait que cette même obligation soit imposée aux propriétaires voulant obtenir un prêt immobilier, seules 50% des maisons en zone à fort risque sont assurées (Kousky et Shabman, 2015). Nous reviendrons plus en détail sur les cas des États-Unis par la suite.

Pour mieux comprendre les défis posés par le risque inondation il faut revenir aux bases du principe d'assurance. Pour qu'un risque soit assurable il faut qu'il soit quantifiable, mutualisable et imprévisible(OCDE, 2016). Le risque inondation ne remplit pas toujours ces trois critères. La question se pose de savoir comment proposer une prime juste, c'est-à-dire qui reflète le risque, abordable et qui permet à l'assureur d'avoir les fonds nécessaires pour indemniser les assurés. Selon le rapport de l'OCDE trois facteurs principaux posent un problème lors de la tarification d'une assurance inondation.

Table 3.2. **Estimates of the share of properties at high-risk of flooding**

Country	Estimate
Australia	Riverine flooding: 7% of domestic houses[1] 1-in-100 year flooding: 160 000 homes[2]
Austria	Flooding (1-in-30): 150 000 exposed people[3] Flooding (1-in-100): 350 000 exposed people[3] Flooding (1-in-300): 650 000 exposed people[3]
Canada	Flooding (1-in-75): 13% of residential properties[3]
Czech Republic	Flooding (1-in-50): 9-10% of households[3]
Estonia	Flooding (1-in-50): 6 708 residents[3] Flooding (1-in-100): 9 171 residents[3]
Germany	Flooding (1-in-50 to 1-in-200): 7.9% of households[4] Flooding (1-in-50 or higher): 1.9% of households[4]
Ireland	Flooding: 300 communities identified as facing significant risk of damaging floods (based on index of hazard and consequences)[5]
Italy	Flooding and landslide (high-risk): 1.1 million residential buildings (9% of total)[9]
Latvia	Flooding (1-in-75): <1%[3]
Portugal	2% of mainland Portugal displays high or very high vulnerability[6]
Russia	7 400 settlements are located in "flood hazard areas"[3]
Spain	Flooding (1-in-100): 3.3% of population[3]
United Kingdom	Some degree of flood risk: 6 million properties (16.7%)[7] Riverine and coastal flooding (1-in-75): 560 000 properties (England and Wales)[7]
United States	Riverine flooding (1-in-100): 4.9 million housing units[8] Coastal flooding (1-in-100): 3.8 million housing units[8] Coastal flooding (1-in-100): 16.4 million residents (5% of population)[3]

Sources: [1] Allianz Australia Insurance Ltd. (2011); [2] Collins and Simpson (2007); [3] Country responses to an OECD questionnaire on the financial management of flood risk (2015); [4] GDV (2015a); [5] Office of Public Works (2012); [6] Costa et al. (2014); [7] Ramsbottom, Sayers and Panzeri (2012); [8] National Research Council (2015); [9] Swiss Re (2015b).

FIGURE 3: Part des habitations exposées par pays, OCDE, 2016

Le premier est la taille des pertes possibles. La fréquence élevée et la sévérité importante dans les zones à risque expliquent les primes élevées et rédhibitoires. Les coûts liés aux inondations sont supérieurs aux autres catastrophes naturelles :en Australie, le montant moyen des sinistres liées aux inondations du Queensland était de 45 374 AUD, alors que le montant des sinistres liés au Cyclone Yasi était de 15 959 AUD (OCDE, 2016).

Le second est la diversification du risque. En effet, la mutualisation n'est possible qu'en ayant un portefeuille diversifié couvrant un large territoire afin de repartir la charge sinistre entre l'ensemble des assurés. Si les risques sont corrélés et qu'il n'y a pas assez de clients faiblement exposés au risque, il s'en suit une hausse des primes. Il est difficile d'attirer des clients à bas risque dans un portefeuille inondation. Les propriétaires ne vivant pas près d'un cours d'eau, n'auront pas tendance à souscrire à une assurance. Seuls les propriétaires vivant dans des zones à risques souhaiteront souscrire, ne laissant dans le portefeuille que les clients fortement exposés. Comme nous le montre la figure 3, la part des propriétés fortement exposées est faible. Ce qui renforce la sous-estimation du risque des clients faiblement exposés et empêche la constitution d'un portefeuille mutualisé avec des "bons" et "mauvais" clients. C'est ainsi que lorsque l'assurance inondation n'est pas inclue *de facto* dans une autre police mais qu'elle est optionnelle alors les clients à faible risque ne souhaiteront pas souscrire. Aux États-Unis, seuls 1 % des propriétaires vivant dans des zones à risque de crue cinq-centennale sont assurés.

La troisième est l'incertitude quant à la modélisation de la charge sinistre. La faible fréquence des catastrophes naturelles et la forte variabilité des dégâts occasionnés, comparée aux autres risques, entraînent des difficultés pour disposer d'un outil statistique robuste pour modéliser l'exposition de l'assureur. Les inondations peuvent provenir de diverses sources : crues, crues éclaires, inondation côtières, inondations urbaines. Cela nécessite une grande variété de modélisation, non seulement dans les zones proches des cours d'eau, mais aussi dans les zones éloignées soumises au risque d'inondation par fortes pluies. De plus, les sinistres occasionnés sur les biens et propriétés varient fortement. Les conditions topographiques des terrains, la capacité d'absorption des sols, la capacité de drainage des systèmes de protection, les différences d'élévation entre les bâtiments entraînent une forte variabilité entre chaque assuré sinistré.

Un des principaux facteurs de la faible demande en assurance inondation, en plus du prix élevé, est la sous-estimation du risque. Les événements à faible probabilité d'occurrence ont tendance à être sous-estimés lorsque les assurés n'ont pas subi de perte (McClelland, Schulze et Coursey, 1993). Cette expérience passée doit être prise en considération pour comprendre les biais auxquels font face les propriétaires : le biais de myopie : la tendance à se focaliser uniquement sur une période à court-terme pour décider du choix de mesure de protection ; le biais d'amnésie : l'oubli rapide des catastrophes passées ; le biais d'optimisme : la tendance à sous-estimer le risque de sinistre ; le biais d'inertie : la tendance à ne rien faire lorsque l'on fait face à de l'incertitude ; le biais de mimétisme : le tendance à fonder ses choix en fonction de celui des autres. Ces biais affectent la perception du risque et conduisent les propriétaires, notamment les moins exposés au risque, à ne pas s'assurer (Kunreuther, 2018).

C'est finalement un cercle vicieux qui entraîne l'échec d'une assurance inondation viable : la faible demande d'assurance entraîne un portefeuille mal diversifié, cela entraîne des primes élevées qui vont encore plus freiner la demande.

2 Chapitre II : Risque inondation aux États-Unis

Notre base de données étant américaine, nous proposons dans cette partie une analyse du risque inondation aux États-Unis et la manière dont les autorités publiques gèrent ce risque.

2.1 Causes

En plus de son littoral, les États-Unis possèdent d'importantes ressources hydrologiques. Avec une superficie de 3 238 000 km², le bassin du Mississippi (Figure 4) est le plus grand bassin d'Amérique du Nord et le troisième au monde après celui de l'Amazone et du Congo. Le fleuve du Mississippi, long de 3 780 km, traverse 10 états : le Minnesota, le Wisconsin, l'Iowa, l'Illinois, le Missouri, le Kentucky, le Tennessee, l'Arkansas, le Mississippi ainsi que la Louisiane. Son bassin versant couvre 41 % du territoire et près d'un Américain sur quatre y vit.

Figure 4: Bassin du Mississippi

En 1993, des inondations dévastatrices ont touché le Midwest américain dans le bassin du Mississippi. Ce sont les inondations les plus coûteuses de l'histoire des États-Unis avec 15 milliards $ de dégâts. D'importante précipitations de juin à août sur des sols déjà humides ont entraîné une hausse du niveau des rivières, d'abord au niveau dans les rivières du Wisconsin et du Minnesota avant d'atteindre le fleuve du Mississippi. La station de mesure du fleuve Mississippi à Saint-Louis atteint son plus haut niveau historique le 01/08/1993 avec 49.58 pieds. (Johnson et al., 2003)

2.2 Impact économique

Table 1. Top 20 Significant Flood Events Covered by the National Flood Insurance Program
(1978-November 30, 2012; $ nominal)

Rank	Event	Date	Number of Paid Losses	Amount Paid	Average Paid Loss
1	Hurricane Katrina	Aug. 2005	167,671	$16,264,188,476	$97,001
2	Hurricane Ike	Sept. 2008	46,412	2,664,167,040	57,391
3	Hurricane Ivan	Sept. 2004	27,658	1,590,436,206	57,504
4	Hurricane Irene	Aug. 2011	43,848	1,302,111,631	29,696
5	Tropical Storm Allison	June 2001	30,663	1,103,877,235	36,000
6	Louisiana Flood	May 1995	31,343	585,071,593	18,667
7	Hurricane Isabel	Sept. 2003	19,869	493,452,308	24,835
8	Hurricane Rita	Sept. 2005	9,517	472,774,099	49,677
9	Hurricane Floyd	Sept. 1999	20,437	462,252,753	22,618
10	Tropical Storm Lee	Sept. 2011	9,748	442,259,918	45,369
11	Hurricane Opal	Oct. 1995	10,343	405,527,543	39,208
12	Tropical Storm Isaac	Aug. 2012	10,126	407,251,178	40,218
13	Hurricane Hugo	Sept. 1989	12,840	376,433,739	29,317
14	Hurricane Wilma	Oct. 2005	9,614	365,030,822	37,975
15	Nor'Easter	Dec. 1992	25,142	346,150,356	13,768
16	Midwest Flood	June 1993	10,472	272,819,515	26,052
17	PA, NJ, NY Floods	June 2006	6,423	228,743,070	35,613
18	Torrential Rain – TN	Apr. 2010	4,108	228,248,545	55,562
19	Nor'Easter	Apr. 2007	8,636	225,657,504	26,130
20	Hurricane Fran	Sept. 1996	10,315	217,843,972	21,119

Source: U.S. Department of Homeland Security, Federal Emergency Management Agency, *Significant Flood Events as of November 30, 2012*, located at http://www.fema.gov/policy-claim-statistics-flood-insurance/policy-claim-statistics-flood-insurance/policy-claim-13-9.

FIGURE 5: 20 événements les plus coûteux de l'histoire du National Flood Insurance Program, King, *Report for Congress*,2013

Ce tableau (Figure 5) nous montre les 20 événements ayant entraîné les plus grands montants de sinistres. L'ouragan Katrina a été une des catastrophes naturelles les plus dramatiques de l'histoire des États-Unis. Le coût total des pertes est estimé à

160 milliards $. Sur ces 160 milliards $, 16.3 milliards ont été payés par le National Flood Insurance Program.

2.3 Prévention et couverture de la population : National Flood Insurance Program

2.3.1 Histoire et problématiques

Au début du XXè siècle, l'assurance inondation était proposée par des assureurs privés. Les inondations du fleuve Mississippi en 1927 et d'autres inondations l'année suivante ont contraint les assureurs à se retirer du marché, arguant que ce risque était inassurable. L'unique aide disponible pour les sinistrés était l'aide fédérale d'urgence. C'est en 1951 que le président Truman propose au Congrès d'établir un véritable système d'assurance inondation, fondé sur une couverture privée avec l'État fédéral qui intervient en tant que réassureur. Le but était de proposer une couverture abordable en limitant la garantie, créant un portefeuille varié géographiquement et en proposant une réassurance publique (NRC, 2017). Malgré cette proposition, aucun système n'est mis en place.

En 1955, après la saison des ouragans, le Président Eisenhower souhaite un système où les propriétaires, les États et le gouvernement fédéral partagent la charge du coût engendré par les inondations. Le Federal Flood Insurance Act en 1956 crée le Federal Flood Indemnity Administration une assurance inondation, un programme de réassurance ainsi qu'un programme de prêt mais, par faute de moyens financiers alloué par le Congrès, cet entité est dissolue en juillet 1957. En 1965, le *Southest Hurricane Disaster Relief Act* est voté au congrès afin de venir en aide aux victimes de l'ouragan Betsy, qui frappa le sud des États-Unis en août 1965 et causa 1,42 milliards $ de dégâts. La répétition des programmes d'aides du gouvernement aux victimes de catastrophes naturelles sur une courte période incite le congrès a demandé une étude d'une autre forme d'aide pérenne pour les sinistrés. En août 1966, le Président Johnson transmet un rapport au Congrès dans lequel est mentionné la volonté d'établir un programme d'assurance réalisable afin d'aider les propriétaires à gérer le risque d'inondation et à décourager l'installation en zone inondable. Il est mentionné que les propriétaires souhaitant s'installer dans des zones à risques devaient être conscients du risque et supporter le coût d'une assurance entièrement liée au risque.

C'est en 1968 qu'est créé le National Flood Insurance Program. Ses deux objectifs sont, premièrement, d'inciter les communautés locales d'utiliser des outils d'aménagement du territoire afin de limiter le développement de construction en zone inondable et, deuxièmement, de fournir une assurance inondation grâce à un partenariat public-privé (NRC, 2017). Cette assurance n'est disponible qu'aux habitants vivant dans des villes ayant accepter de limiter le développement en zone inondable. Un des principes du NFIP est de proposer des primes abordables tout en

laissant les assureurs privés supporter une partie du risque. Le NFIP peut emprunter auprès du Trésor public afin d'indemniser les sinistrés lors des années sans grandes catastrophes. Le Trésor public intervient en tant que réassureur et supporte la charge liée aux grandes catastrophes entraînant d'important dégâts. Ainsi les primes demandées par le NFIP n'intègrent pas la partie liée aux risques de catastrophes naturelles car ce risque est supporté par le Trésor et n'atteignent pas de niveaux rédhibitoires pour les propriétaires.

De plus, certaines propriétés payeront une prime sous-tarifée : les propriétaires de maisons située en zone à risque avant l'introduction de carte délimitant les zones à risques se verraient proposer une prime beaucoup trop élevée. Une distinction est faite entre les habitations construites avant la délimitation des zones à risque et celles construites après : seules ces dernières paieront une prime reflétant entièrement le risque en souscrivant auprès des assureurs privés. L'idée sous-jacente était que ces propriétés allaient être de moins en moins nombreuses dans le portefeuille en raison des destructions provoquées par les inondations (NRC, 2017).

Malgré cette proposition, l'idée d'un partage des risques avec des assureurs privés n'a pas été retenue. Seul le NFIP est responsable de la tarification et gère le risque. Les assureurs privés n'interviennent que dans la commercialisation des produits. L'éligibilité à la protection repose sur l'adoption d'aménagement du territoire et la production de FIRM (*Flood Insurance Rate Map*), à savoir des délimitations géographiques des zones à risques : ces zones à risques sont celles qui présentent un risque d'être affectées par une crue centennale, appelées *Special Flood Hazard Areas*.

L'année 2005, marquée notamment par l'ouragan Katrina, a été l'année la plus coûteuse pour le NFIP d'après notre base de données. Au mois d'août 2005 le NFIP a dépensé plus de 16 milliards $ pour indemniser les sinistrés.

Cette année catastrophique a lourdement gérer un énorme déficit pour le NFIP comme on peut le voir sur la Figure 6.

Afin de renforcer la solvabilité du programme, le Congrès vote le *Biggert-Waters Flood Insurance Reform Act*. Cette réforme vise à augmenter les primes payées par les propriétaires afin qu'elles reflètent au mieux le risque et se rapprochent de la prime actuarielle. La partie subventionnée des primes des propriétaires de maisons secondaires, de locaux professionnels et ceux qui ont été sinistrés à de multiples reprises seront supprimées. Les primes des propriétaires vivant dans des habitations construites avant l'établissement d'une carte des zones inondables seront également revues à la hausse.

Cette hausse des primes, pouvant aller jusqu'à une multiplication par 10 dans certains endroits (Reuters, 2014), a été l'objet de critiques. Faisant face au fait que de nombreux habitants ne pourraient plus payer leur primes, le Congrès a voté le Homeowner Flood Insurance Affordability Act en 2014. Cette loi retarde l'entrée en vigueur des hausses de primes prévues par l'ancienne réforme et les supprime dans certains cas.

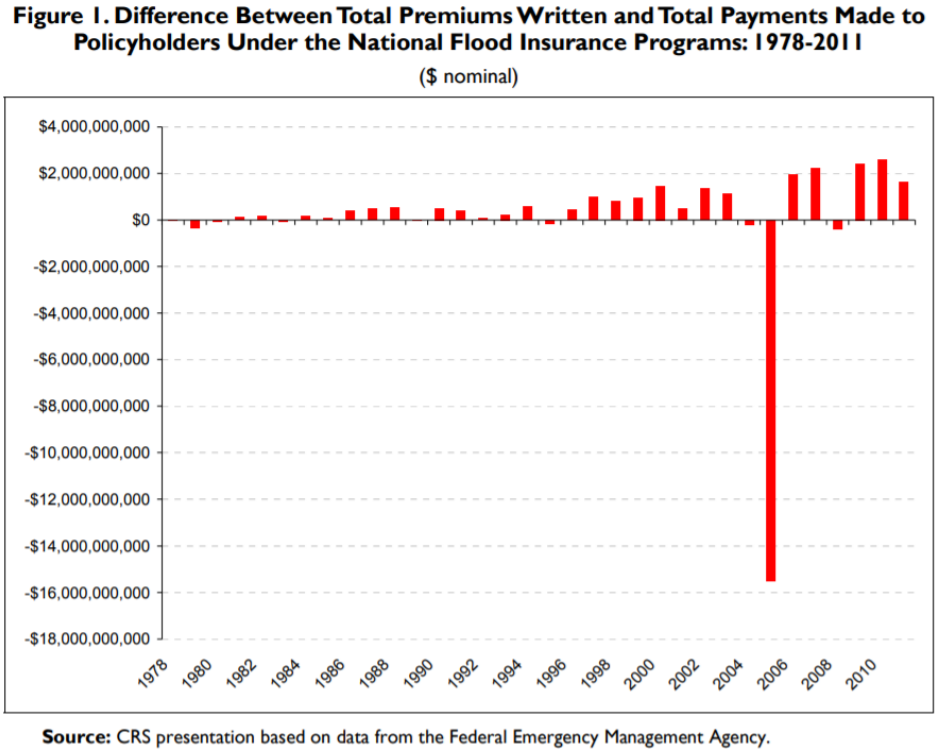


FIGURE 6: Différence entre les primes et le coût des sinistres par année, King, *Report for Congress.* 2013

2.3.2 Cartographie des zones à risque

Définitions

Special Flood Hazard Area (SFHA : Zone ayant une probabilité d'être inondée égale à au moins 1 %.

Base Flood Elevation (BFE) : Hauteur d'eau atteinte lors de la crue centennale.

Zone	Description
A	Zone à risque : Zone à risque (SFHA) sans établissement du BFE
A1-A30,AE	Zone à risque (SFHA) avec établissement du BFE
AO	Zone à risque (SFHA) : inondations peu profondes (1 à 3 pieds)
A99	Zone à risque (SFHA) avec des protections efficientes (Digues/Barrages)
AH	Zone à risque (SFHA) : inondation peu profondes avec établissement du BFE
AR	Zone à risque (SFHA) : Protections en attente de certification
V	Zone à risque (SFHA) : Inondations par submersion marines
V1-V30, VE	Zone à risque (SFHA) : Inondations par submersion marines avec établissement du BFE
B, X	Zone à risque modéré
C, X	Zone à risque faible
D	Zone à risque indéterminé mais possible

Table 1: Classification des zones inondables

Au sein des zones A et V, certains territoires possèdent des informations supplémentaires sur la structure topographique du terrain. Ces zones sont nommées AE et VE. Les zones AE, étaient anciennement des zones numérotées de A01 à A30. Les zones AH sont des zones dont la hauteur d'eau atteinte lors d'inondations est peu élevée dans des zones sans relief.

Pour les habitations en zone A non numérotée, trois types de tarification s'appliquent : Un qui utilise l'élévation de la maison par rapport à la hauteur d'eau de la crue centennale lorsque celle-ci a été certifiée. Un second qui utilise la niveau d'élévation du sol de la maison. Enfin, un dernier pour les habitations où les détails sur le niveau d'élévation de la structure n'est pas disponible.

2.3.3 Tarification

Primes reflétant le risque :

Le NFIP considère que les primes fondées sur le risque sont celles qui permettent de couvrir les sinistres attendus du portefeuille. Au sein des zones à risque (zone A et V), des modèles hydrologiques sont utilisés pour déterminer les primes. En dehors de ces zones à risque (zone X), le passé sinistre est utilisé pour les calculer : dans les zones à faible risque il ne serait pas rentable de mener des études hydrologiques poussées afin de déterminer les primes. (Kousky et Shabman, 2014)

Le NFIP établit les primes à payer en fonction du montant de la valeur assurée des biens et de la structure. En fonction de certaines caractéristiques, deux taux sont déterminés : un à appliquer sur la valeur assurée des biens et l'autre sur la valeur assurée de la structure. La somme des deux montants permet d'établir la prime pure.

$$Taux = \left[\sum_{i=Min}^{Max} (PELV_i \times DELV_i) \right] \times \frac{LADJ \times DED \times UINS}{EXLOSS}$$

avec

- $PELV_i$= Probabilité que l'eau atteigne une hauteur i
- $DELV_i$= Pourcentage de destruction relative à une hauteur d'eau i
- $LADJ$ = Facteur de chargement lié aux frais de gestion
- DED= Facteur lié à la franchise
- $UINS$= Facteur correspondant à la sous-assurance
- $EXLOSS$= Facteur lié au Loss-Ratio attendu incluant les frais restants

Probabilité d'élévation

Les formules utilisées par le NFIP utilise la probabilité d'élévation de l'eau relative à la hauteur d'eau attendue lors de la crue centennale. Les probabilités sont calculées de la manière suivante :

$$-log_{10}(P(elev)) = C_1 + C_2 elev + C_3 elev^2 + C_4 elev^3 + C_5 elev^4$$

avec

- C_x : Les coefficients relatifs à la zone topographique
- $elev$: Elevation relative à la hauteur de la crue centennale
- $P(elev)$: Probabilité que l'eau atteigne le niveau $elev$

Le risque auquel sont soumis les habitations dépend de l'élévation de la maison et la structure de celle-ci, mais aussi de la structure topographique du terrain. Les zones topographiques sont numérotées de A01 à A30. Elles sont caractérisées par la différence d'élévation entre la crue décennale et la crue centennale. La zone A01 correspond à une zone avec un faible relief, la zone A30 correspond au contraire à une vallée aux pentes abruptes. Le NFIP, cependant, ne différencie pas les primes selon la zone topographique. Les courbes de probabilité d'élévation sont calculées à partir d'une moyenne pondérée de chacune des courbes des zones topographiques. Six courbes sont utilisées, correspondant à six types de topographies différentes. La pondération est déterminée, en analysant les propriétés assurées soumises à un risque de crue décennale. Si 10 % des habitations assurées, pour lesquelles il y a des informations détaillées, dans un Etat sont soumises à un risque de crue décennale dont la hauteur d'eau sera de -4 ft par rapport à la crue centennale, alors le NFIP supposera que 10 % des habitations de l'État sont soumises à ce même risque. En prenant ensuite la part d'assurés pour chaque État, le NFIP crée une pondération au niveau national.

Zones	Différence d'élévation entre la crue décennale et centennale	C_1	C_2	C_3	C_4	C_5	Profondeur minimale	Profondeur maximale
A01 à A02	0 à 1.25 ft	1.99028	2.94155	2.45622	1.13909	0.201998	-2.2	0.2
A03 à A06	1.25 à 3.25 ft	1.99808	0.531039	0.0287727	0.008155350	0.00177257	-4.0	1.2
A07 à A10	3.25 à 5.25 ft	1.99091	0.236907	0.00714333	0.001169790	0.000101835	-6.4	2.6
A11 à A13	5.25 à 6.75 ft	1.99494	0.181604	0.00471243	0.000159033	-0.000028376	-8.0	3.5
A14 à A17	6.75 à 8.75 ft	1.99384	0.143237	0.00391510	0.000623394	0.0000386951	-9.5	4.0
A18 à A30	8.75 ft et au-dessus	1.99669	0.108811	0.00483903	0.000650387	0.0000251549	-13.0	5.0

Table 2: Paramétrage du calcul de la probabilité d'élévation

Zones	Différence d'élévation entre la crue décennale et centennale	C_1	C_2	C_3	C_4	C_5	Profondeur minimale	Profondeur maximale
A01 à A02	0 à 1.25 ft	1.78915	2.365770	1.77562500	0.783071	0.135608	-2.2	0.2
A03 à A06	1.25 à 3.25 ft	1.80047	0.427384	0.006967	0.002012	0.000727	-4.0	1.2
A07 à A10	3.25 à 5.25 ft	1.79699	0.190719	0.002094	0.000267	0.000013	-6.4	2.6
A11 à A13	5.25 à 6.75 ft	1.79437	0.142739	0.002329	0.000165	-0.000021	-8.0	3.5
A14 à A17	6.75 à 8.75 ft	1.79925	0.112159	0.001721	0.000513	0.000037	-9.5	4.0
A18 à A30	8.75 ft et au-dessus	1.79239	0.082063	0.003351	0.000632	0.000027	-13.0	5.0

Table 3: Paramétrage pondéré du calcul de la probabilité d'élévation

Afin de prendre en compte le biais lié au manque de profondeur historique des données à certains endroits, (20 % des communes selon le NFIP), d'autres courbes sont utilisées, appelées PELV 500. Une moyenne pondérée (20 % des coefficients du PELV500 et 80% des coefficients du PELV) est utilisée pour déterminer les courbes finales. La FEMA considère par exemple que dans une zone où un historique de 25 années est disponible, la probabilité de la crue centennale calculée ne serait pas de 1 % mais de 1.59 %. Le risque est donc sous-estimé. On introduit une surestimation artificielle.

Afin d'homogénéiser les taux et d'avoir un unique taux pour les zones AE, le NFIP calcule une moyenne pondérée selon la différence d'élévation entre la structure et la hauteur d'eau de la crue centennale.

Cette prise en compte des biais liées aux manque de données peut être problématique (Congresssional Budget Office, Congress of the United States, 2009) . Elle

génère une subvention croisée entre tous les assurées habitant en zone AE et pas uniquement ceux habitant dans les zones à faibles observations. De plus, cela signifie que la délimitation de la zone AE n'est pas aussi large qu'elle ne devrait l'être et que certaines habitations vivant en zone X (à faible risque) sont en fait menacées par la crue centennale. Cependant, la tarification en zone X se fait sur l'historique des sinistres sur toutes les zones X, ainsi une exposition plus importante au risque d'inondation se reflétera dans cet historique et donc dans les primes ajustées.

Probabilité de dommage

Chaque année, une courbe de dommage est établie. Elle caractérise le pourcentage de destruction liée à la différence entre la hauteur d'eau de l'inondation et le plus bas niveau de la structure. Pour déterminer chaque courbe de destruction, la FEMA utilise les données calculées par le Corps militaires des ingénieurs et son passé sinistre, en enlevant l'année 2005 considérée comme exceptionnelle.(Congresssional Budget Office, Congress of the United States, 2009)

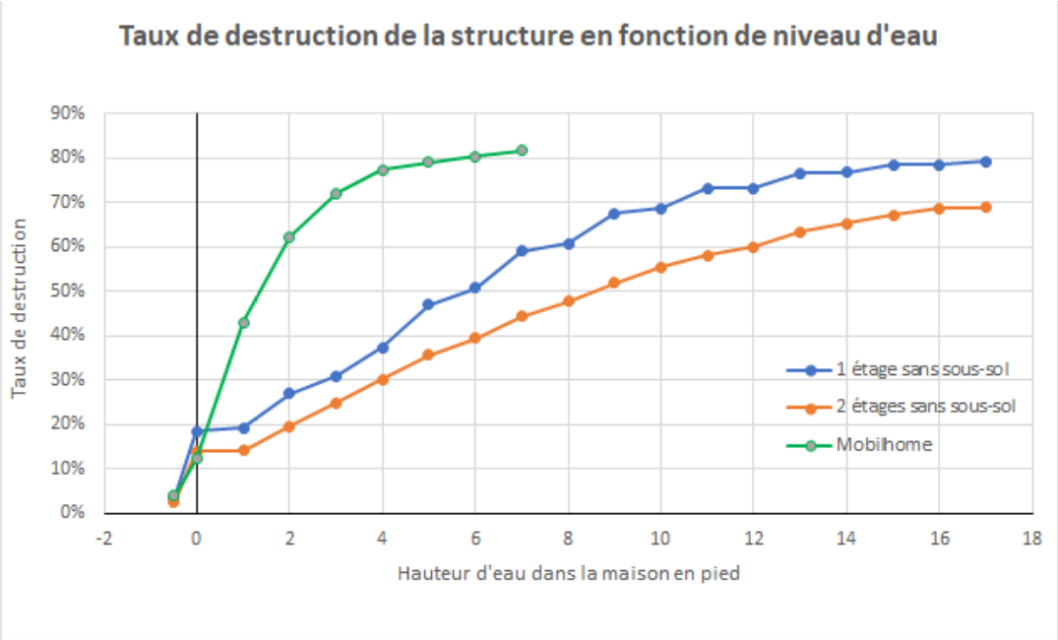


FIGURE 7: Taux de destruction selon le type de la structure

Preferred risk policy

Les PRP sont des polices d'assurance applicables pour les assurés vivant en zone X et qui ont un faible passé sinistre : les propriétés qui remplissent toutes les conditions suivantes sont éligibles à ces tarifs à bas coût :

- Deux sinistres de plus de 1000 $ chacun.
- Trois sinistres ou plus.
- Deux aides fédérales d'urgence de plus de 1000 $ chacune.

— Trois aides fédérales d'urgence ou plus.
— Un sinistre et deux aides fédérales d'urgence.

Primes subventionnées

Trois types de subventions sont mises en place au sein du programme d'assurance :
— Un pour les maisons construites en zone à risque avant l'établissement d'une cartographie du risque, ce que l'on appelle les *pre-firm properties*.
— Un pour les maisons vivant dans des villes ayant prises des mesures de prévention : *Community Rating System Discounts*.
— Un pour les maisons dont l'actualisation de la cartographie les a faites passées d'une zone à faible risque à une zone à fort risque : *Grandfathered Policies*.

Subventions des habitations Pre-FIRM

Les habitations construites avant l'établissement d'une cartographie des zones inondables, et se trouvant dans une zone à risque, ont une réduction sur leur prime. Celle-ci ne s'applique que sur le montant minimum de la valeur assurée (60 000 $). La prime de ces habitations n'est pas calculée en fonction de l'élévation de l'habitation par rapport au niveau de la crue centennale comme pour les propriétés en zone à risque. Ces subventions ont été mises en place afin d'encourager les propriétaires à s'assurer.

Nous avons vu, au début de notre partie, que lorsqu'il y avait la présence d'assureurs privés dans le pool, le Trésor public les indemnisait lorsqu'ils établissaient des primes qui ne reflétaient pas le risque des propriétés construites avant l'établissement de la cartographie des zones inondables. Le départ des assureurs privés du pool a mis un terme à ces versements. Les subventions accordées sont calculées à partir du principe suivant : les primes subventionnées des polices pre-FIRM et celles qui ne sont pas subventionnées doivent permettre de couvrir la moyenne annuelle de la charge sinistre depuis le début du programme (Kousky et Shabman, 2014). Cela entraîne donc une subvention croisée (*National Research Council*, 2015) mais qui n'est pas parfaite car cette moyenne de charge sinistre n'intègre pas les années catastrophiques.

Community Rating System Discounts

Les habitants vivant dans des villes qui mettent en œuvre des mesures de prévention au-delà de celles imposées par le NFIP peuvent obtenir des réductions sur leur prime. Ces réductions vont de 5% à 45% et s'appliquent aussi bien dans les zones inondables que dans les zones à faible risque avec cependant une différence de taux selon les zones. Un classement est établi avec 9 classes : les communautés en classe 9 ont une réduction de 5 % et les habitants des communautés en classe 1 ont une réduction de 45%. Seules 7% des villes éligibles au programme du NFIP font

partie du système CRS. Cependant ces communautés comptent pour 70% des polices totales (FEMA, CRS Fact Sheet). Afin de garantir la neutralité actuarielle du portefeuille, et parce que les mesures éligibles au CRS n'entraînent pas forcément une baisse directe de la sinistralité, l'ensemble des primes sont chargées (13.4 %). Il y a donc subvention croisée entre les collectivités qui participent au programme et les autres.

Grandfathered Policies

Les *Grandfathered policies* sont des polices d'assurances disposant d'une clause d'antériorité : par exemple si une mise à jour de la cartographie des zones inondables fait passer une propriété d'une zone à faible risque à une zone fort risque, alors la police garde sa tarification avantageuse liée à son ancienne zone. Ces polices ne sont cependant pas éligibles au *Preferred risk policy*. C'est donc une prime standard, celle qui s'applique aux habitations en zone X, qui est payée par les assurés. Un autre type de clause d'antériorité est celle lié à une modification de la hauteur d'eau de la crue centennale. Dans ce cas il n'y a pas d'augmentation du tarif et l'élévation antérieure est conservée dans le calcul de celui-ci. Afin de compenser ces tarifs avantageux, le NFIP applique un chargement supplémentaire sur le tarif des assurés vivant en zone inondable. C'est un autre exemple de subvention-croisée au sein du portefeuille. D'après un rapport du congrès américain (*Congressional Research Service*), 9 % des polices en 2018 comportaient cette clause.

3 Chapitre III : Analyse du portefeuille

3.1 Introduction

Nos données proviennent de la FEMA (Federal Emergency Management Agency), l'agence américaine qui est en charge du NFIP. Nous disposons de l'historique des sinistres couverts par le NFIP de 1973 à 2019, soit plus de deux millions de lignes. Les variables explicatives qui nous intéressent sont :

Basement/Enclosure/Crawlspace type	Type de sous-sol répertorié en 5 catégories : 0 - Aucun 1 - Sous-Sol fini 2 - Sous-Sol non fini 3 - Vide sanitaire 4 - Vide sanitaire en sous-sol
Reported City	Ville
CountyCode	Code du Comté
Dateofloss	Jour où l'eau a atteint l'habitation
floodZone	A - Zone à fort risque, sans BFE sur la FIRM AE - Zone à fort risque, avec BFE sur la FIRM X - Zone à risque faible ou modérée
Numberoffloorsintheinsuredbuilding	Code qui indique le nombre d'étages de la maison
OriginalConstructionDate	Date de construction de la maison
AmountPaidOnBuildingClaim	Montant du dommage sur la structure de la maison
AmountPaidOnContentsClaim	Montant du dommage sur le contenu de la maison
PostFIRMConstructionIndicator	Maison construite après l'établissement de la FIRM
PrimaryResidence	Résidence primaire
CommunityRatingSystemDiscount	Taux de réduction sur la prime payée en fonction des mesures de prévention adoptées par la commune de l'assuré
elevatedBuildingIndicator	Indicateur si la maison est conforme aux standards d'élévation du NFIP
LocationofContents	Endroits où se trouve le contenu assuré 1 - Sous-sol seulement 2 - Sous-sol et niveau supérieur 3 - Premier niveau de la maison au dessus du sol seulement 4 - Premier niveau au dessus du sol et niveaux supérieurs

Table 4: Liste des variables explicatives

On crée la variable de taux de destruction tel que

$$Taux\,de\,destruction = \frac{Dommage\ structure + Dommage\ contenus}{Valeur\ assurée\ structure + Valeur\ assurée\ batiment}$$

3.2 Traitement des données

Dans notre base de sinistres nous avons 4668 lignes pour lesquelles il n'a pas de valeur assurée et qui nous empêchent de connaître le taux de destruction.

3.3 Statistiques descriptives

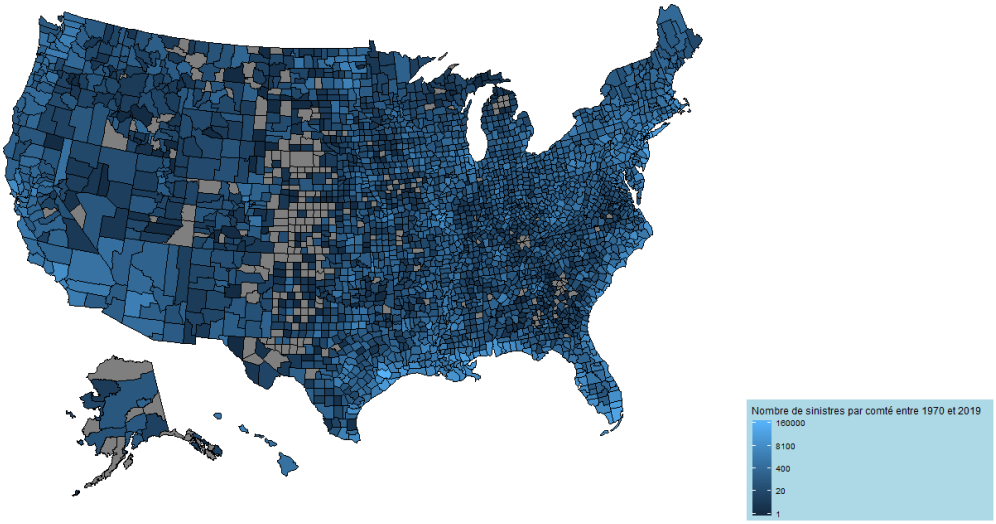


FIGURE 8: Nombre de sinistres par comté entre 1973 et 2019

Le graphique nous montre l'ensemble des sinistres par comté de 1973 à 2019. On observe que les comtés situés sur les littoraux sont les plus sinistrés et notamment le sud-est des États-Unis en raison des ouragans.

Lorsqu'on observe le nombre de sinistres par année, on observe que les années les plus sinistrées ont été 2005, 2012 et 2017 en raison des ouragans, respectivement, Katrina, Sandy et Harvey. On remarque ainsi la forte variabilité de la fréquence des sinistres.

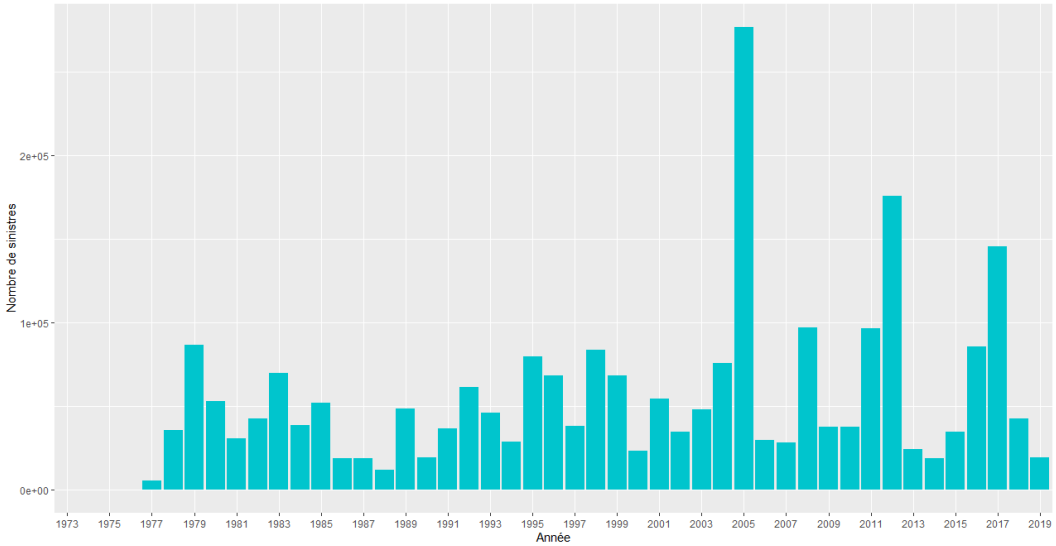


FIGURE 9: Nombre de sinistres par année entre 1973 et 2019

3.4 Modélisation du cout des sinistres

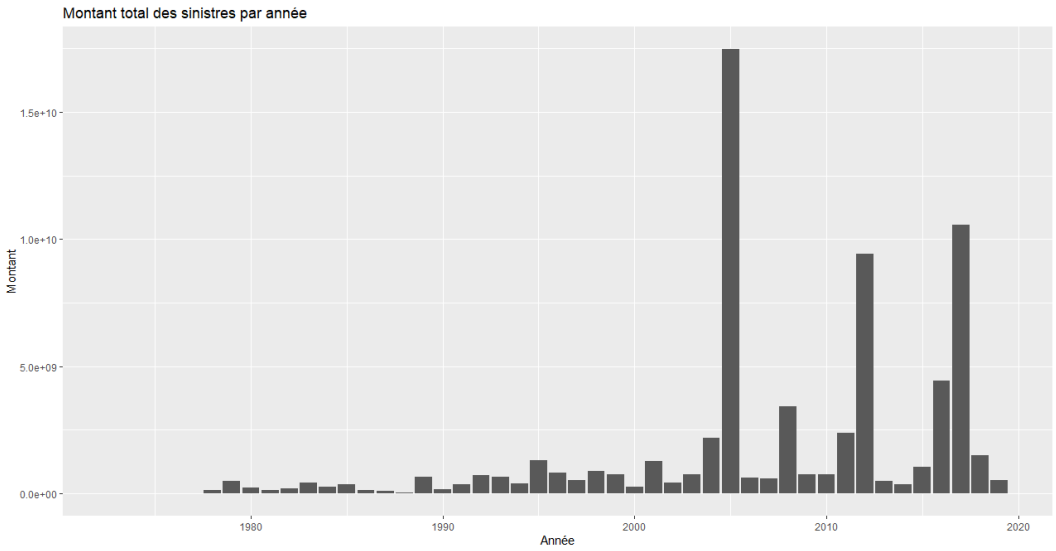


FIGURE 10: Montant des sinistres par année entre 1973 et 2019

Sur les 69 milliards $ payés par le NFIP entre 1973 et 2019, 17 milliards ne concernent que l'année 2005. En 2012, l'année de l'ouragan Sandy, le NFIP a versé près de 9,4 milliards $ et 10,4 milliards $ en 2017. On observe qu'en dehors de ces trois années de sinistralités exceptionnelles, quelques années sont remarquables : 2004, 2008, 2011, 2016. En dehors de ces années, la sinistralité ne dépasse pas 1,3 milliard $.

Sur l'ensemble des 2 397 035 sinistres déclarés entre 1973 et 2019, 550 777 n'ont pas procédé à une indemnisation.

3.5 Modèles linéaires généralisés

3.5.1 Introduction

Le MLG comporte trois composantes :

- La composante aléatoire : les variables à expliquer $Y_1, Y_2, Y_3, ..., Y_n$
- La composante déterministe : les vecteurs explicatifs $X_1 = (X_{11}, X_{12}, X_{13}, ..., X_{1n}), ..., X_p = (X_{p1}, ..., X_{pn})$
- Une fonction lien g qui exprime une relation fonctionnelle entre la composante aléatoire et la composante déterministe. La fonction g est strictement monotone, définie sur $\mathcal{R}$ telle que :

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

On a également : $E(Y) = b'(\theta)$ $V(Y) = b''(\theta)a(\phi)$ où $b''(\theta) = V(\mu)$ est la fonction variance.

3.5.2 Estimation des paramètres

La log-vraisemblance du modèle s'écrit :

$$log\mathcal{L}(\theta_1, ...\theta_n, \phi, y_1, ..., y_n) = \sum_{i=1}^{n} [\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\omega_i + c(y_i, \phi)]$$

Pour obtenir les paramètres β on doit dériver la log-vraisemblance par rapport aux β_i et d'utiliser les conditions du premier ordre :

On note $\mu_i = E(Y_i)$ et $\nu_i = g(\mu_i) = X_i\beta$

$$\frac{\partial ln(\mathcal{L}_i)}{\partial \beta_j} = \frac{\partial ln(\mathcal{L}_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{y_i - \mu_i}{V(Y_i)}X_{ij}$$

Les équations sont donc :

$$\sum_{i=1}^{n} \frac{\partial ln(\mathcal{L}_i)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \eta_i}\frac{y_i - \mu_i}{V(Y_i)}X_{ij} = 0 \quad \forall j \in \{1, ..p\}$$

On ne peut pas résoudre ces équations analytiquement.

Les logiciels statistiques utilisent généralement deux algorithmes pour calculer les estimateurs du maximum de vraisemblance : l'algorithme du score de Fisher et l'algorithme IRLS.

3.5.3 Qualité du modèle

Déviance

On définit le modèle saturé comme le modèle qui a le meilleur ajustement possible. On l'obtient si on a autant d'observations que de paramètres à estimer. C'est un modèle qui reproduit la réalité au lieu de la résumer.

On va comparer notre modèle au modèle saturé grâce aux vraisemblances des deux modèles. Si notre modèle est bon alors sa vraisemblance sera proche de celle du modèle saturé. Soit $\lambda = \frac{\mathcal{L}_{sat}}{\mathcal{L}}$

$ln(\lambda) = l_{sat} - l$ et $\boxed{D = 2ln(\lambda)}$ la déviance réduite

$D \xrightarrow{\mathcal{L}} \mathcal{X}^2_{n-p-1}$

On considère que l'ajustement du modèle est bon si $D/n - p - 1 \approx 1$

On considère qu'il est mauvais quand $D \geq \mathcal{X}^2_{n-p-1,1-\alpha}$

Statistique de Pearson

On définit la statistique de Pearson par :

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(Y_i)}$$

On compare les valeurs observées à leur prévision par le modèle.

X^2 ~ $\mathcal{X}^2_{n-p-1}$

3.5.4 Test d'hypothèses

Test entre modèles emboîtes

$$H_0 : \beta_{H_0} = (\beta_0, .., \beta_p)$$
$$H_1 : \beta_{H_1} = (\beta_0, .., \beta_q)$$

avec $q < p < n$

La statistique du test est

$$\Delta = D_0 - D_1 = 2(ln(\mathcal{L}_{\hat{\beta}_1}(y)) - ln(\mathcal{L}_{\hat{\beta}_0}(y)))$$

Δ~ $\mathcal{X}^2_{p-q}$

On n'accepte pas H_0 quand $\Delta_{obs} \geq \mathcal{X}^2_{p-q,1-\alpha}$

Tester l'hypothèse de nullité des coefficients revient à se demander si l'on peut fusionner certaines modalité avec le niveau de référence.

Généralisation

Dans certains cas on se demande si l'on peut regrouper certaines modalités entre elles.

On va tester l'hypothèse de contrainte linéaire suivante :

$$H_0 : C\beta = r$$

Nous allons considérer 3 tests :
- Test du rapport de vraisemblance
- Test de Wald
- Test du Score

Test du rapport de vraisemblance

Soient $\mathcal{L}_{\hat{\beta}_1}$ et $\mathcal{L}_{\hat{\beta}_0}$ les fonctions de vraisemblance des modèles sans et avec contraintes. On s'intéresse à la différence entre les valeurs de la log-vraisemblance.

$2(ln(\mathcal{L}_{\hat{\beta}_1}(y)) - ln(\mathcal{L}_{\hat{\beta}_0}(y))) \sim \mathcal{X}^2_q$

avec q le nombre de lignes sur la matrice C.

Test de Wald

On s'intéresse à la différence entre l'estimateur du maximum de vraisemblance entre les deux modèles.

$\hat{\beta} \sim \mathcal{N}(\beta, \phi(X^tWX)^{-1})$

avec W une matrice de pondération telle que :

$W = diag[\frac{1}{Var(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i})^2]_{1 \leq i \leq n}$

La statistique de test est :

$(C\hat{\beta} - r)(\phi C(X^tWX)^{-1}C^t)^{-1}(C\hat{\beta} - r) \sim \mathcal{X}^2_q$

Test du score

On s'intéresse à la dérivée de la log-vraisemblance.

$l'(\beta) = \phi^{-1}X^tWG(y - \mu)$ avec G une matrice diagonale d'éléments $g'(\mu_i)$ et W une matrice diagonale d'éléments $[g'(\mu_i)^2V(\mu_i)]^{-1}$.

On peut démontrer que $E(l'(\beta)) = 0$ et $V(l'(\beta)) = E(l'(\beta)[l'(\beta)]^t) = \phi^{-1}X^tWX$

La statistique du score est donnée par :

$l'(\beta_{H_0})^t[Var(l'(\beta))]^{-1}l'(\beta_{H_0}) \sim \mathcal{X}^2_q$

avec $l'(\beta_{H_0}) = \phi^{-1}X^tWG(y - \tilde{\mu})$ et $\tilde{\mu}$ l'EMV de $E(Y)$ évaluée sous H_0

3.5.5 Validation du modèle

Détection des valeurs influentes : Distance de Cook

Comme pour le modèle linéaire gaussien on cherche à mesurer la différence entre $\hat{\beta}$ et $\hat{\beta}_{(i)}$ le vecteur de coefficients estimés dans le modèle privé de la i-ème observation.

$$C_i = \frac{1}{p+1}(\hat{\beta} - \hat{\beta}_{(i)})^t (X^t W X)(\hat{\beta} - \hat{\beta}_{(i)})$$

Si $C_i \geq 1$ l'observation i est anormale. D'autres approches considèrent le seuil $4/n - p - 1$.

Analyse des résidus

Les mesures de validité nous donne une indication globale sur la qualité du modèle. L'analyse des résidus nous permet d'étudier plus précisément les observations qui sont responsables d'une mauvaise qualité du modèle.

Résidus de Pearson

On définit les résidus de Pearson par :

$$r_i^p = \frac{\sqrt{w_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

et les résidus de Pearson normalisés par :

$$r_i^p = \frac{\sqrt{w_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)(1 - h_{ii})}}$$

Les résidus normalisés supérieurs à 2 ou inférieurs à -2 doivent faire l'objet d'une attention particulière.

Résidus de déviance

On considère que chaque observation y_i contribue à la déviance totale à un niveau d_i avec $D = \sum_{i=1}^{n} d_i$

Les résidus de déviance sont définis par :

$$r_i^D = signe(y_i - \mu_i)\sqrt{d_i}$$

avec $D = \sum_{i=1}^{n} (r_i^D)^2$

$$g(E(\mu)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_K$$

3.6 Modélisation du coût des sinistres par GLM

Dans la mesure où nous allons appliquer notre étude spatiale aux États du bassin du Mississippi et afin d'avoir des événements relativement homogènes (ici le risque d'inondation liées aux débordements de rivières/fleuves), nous sélectionnons l'ensemble des sinistres des États de l'Iowa, du Dakota du Nord et du Sud, du Nebraska, Kansas, Oklahoma, Minnesota, Wisconsin, Illinois, Missouri et l'Arkansas. Dans un second temps, nous isolerons les sinistres causés par l'ouragan Katrina en 2005 pour étudier les différences du modèle lorsque l'on utilise des données issus de sinistres extrêmes.

Modèle à distribution Gamma

Soit Y une variable suivant une loi Gamma. Sa densité peut s'écrire :

$$f(y) = \frac{1}{y\Gamma(\phi^{-1}} \left(\frac{y}{\mu\varphi}\right)^{\phi^{-1}} exp\left(-\frac{y}{\mu\varphi}\right), \ \ \forall y \in \mathbb{R}^+$$

La densité de Y peut s'écrire sous la forme exponentielle :

$$f(y) = exp\left[\frac{y/\mu - (-log\mu)}{-\varphi} + \frac{1-\varphi}{\varphi} logy - \frac{log\varphi}{\varphi} - log\Gamma(\varphi^{-1})\right], \ \ \forall y \in \mathbb{R}^+$$

La fonction lien canonique est $\theta = \mu^{-1}$ et $b(\theta) = -log(\mu)$. La fonction variance s'écrit$V(\mu) = \mu^2$. La variance de Y s'écrit $V(Y) = \varphi E(Y)^2$. On obtient ainsi le coefficient de variation φ tel que,

$$\frac{\sqrt{V(Y)}}{E(Y)} = \varphi$$

Application du modèle à distribution Gamma

Nous avons 151 918 sinistres enregistrés ayant donné lieu à une indemnisation sur les États retenus entre 1974 et 2019. Nous retenons dans le modèle les "*Single family residence*" c'est-à-dire les maisons individuelles afin d'avoir des sinistres comparables. Elles représentent 82 % des sinistres sélectionnés.

Après traitement des données manquantes ou mal enregistrées, nous avons 110 251 sinistres dans notre sélection. Nous prenons 85 % des données pour ajuster le modèle. Les 15 % restants nous servent à tester le modèle.

Multicolinéarité

Lorsque des variables explicatives mesurent un même phénomène, on parle de multicolinéarité. Cela pose un problème dans les modèles de régression en raison de l'instabilité générée dans le modèle. En effet, la multicolinéarité peut augmenter la variance des coefficients estimés et en rendre plus complexe l'interprétation. Les coefficients peuvent être déterminés, à tort, comme non-significatifs et ils peuvent être instables si l'on recalibre le modèle sur un autre échantillon.

La mesure de la colinéarité peut se faire grâce aux facteurs d'inflation de la variance. Cette mesure estime dans quelle proportion la variance d'un coefficient est augmentée à cause de la présence d'une relation linéaire avec d'autres variables explicatives. Par exemple, un VIF de 1.5 désigne le fait que la variance du coefficient est 50% supérieure à ce qui aurait du être estimé si la variable n'était pas corrélée aux autres variables explicatives. En estimant notre modèle Gamma et en déterminant les VIF nous obtenons :

	GVIF	Df	GVIF^(1/(2*Df))
CRS	1.050	1	1.025
SousSol	6.995	4	1.275
PostFIRM	19.548	2	2.103
IndicateurElevation	20.249	2	2.121
EndroitContenu	34.081	6	1.342
ZoneIno	1.219	2	1.051
Etage	31.971	4	1.542
ResidencePrim	1.066	1	1.032

Table 5: Analyse de la multicolinéarité des variables explicatives

Nous souhaitons obtenir un VIF proche de 1 pour considérer l'absence de colinéarité. Nous observons que certaines variables ont un VIF très supérieur à 1. En couplant ces résultats à une matrice de corrélation, on observe que les variables "Etage" et "SousSol" sont très corrélées, ainsi que les variables "PostFIRM" et "IndicateurElevation". Nous estimons à nouveau notre modèle en supprimant les variables "SousSol" et "PostFirm".

	GVIF	Df	GVIF^(1/(2*Df))
CRS	1.041	1	1.020
ZoneIno	1.050	2	1.012
Etage	1.029	4	1.004
ResidencePrim	1.010	1	1.005

Nous n'avons plus de problème de multicolinéarité dans notre modèle.

Dans un premier temps, nous estimons le modèle à partir de notre base d'apprentissage. Nous avons choisis une fonction lien log et considérer la variable" Valeur Assurée" en tant que variable offset. La fonction lien log nous permet d'obtenir un effet multiplicatif et une interprétation plus simple des résultats.

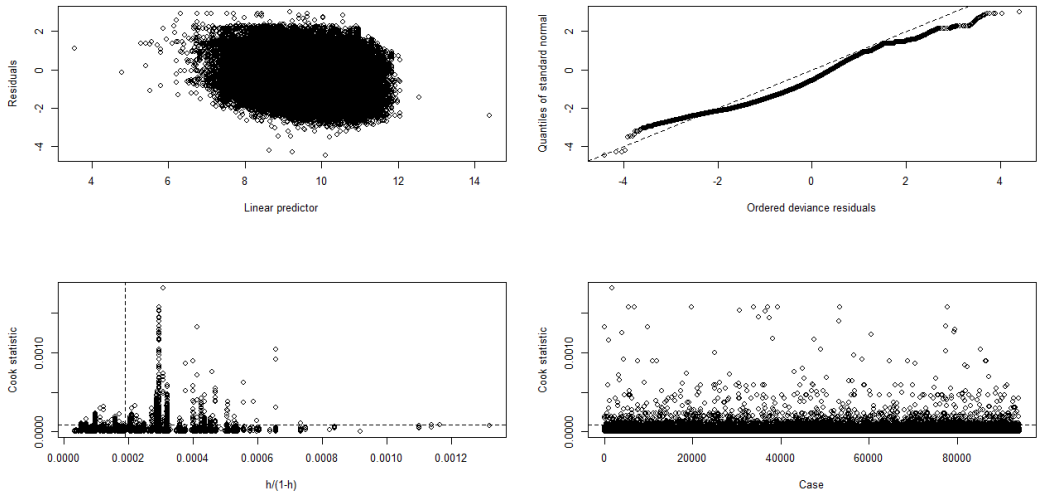


FIGURE 11: Graphiques de diagnostic du modèle Gamma

L'étude des résidus nous permet de voir qu'il y a une certaine structure sur le graphique Résidus vs Valeurs prédites, bien que les résidus soient répartis plutôt aléatoirement entre -2 et 2. Les résidus de déviance ne sont pas alignés avec précision sur la droite d'Henry, sans pour autant s'en éloigner. Nous observons que beaucoup de variables ont une grande influence sur le modèle.

Nous testons notre modèle sur l'échantillon de test.

Nous avons ensuite effectué une validation du modèle sur l'échantillon de test.

	Moyenne	Somme totale
Prédictions	22 313	360 814 345
Valeurs observées	16 969	274 391 537

On remarque que notre modèle surestime les montants de sinistre réels, de plus de 30% sur la somme totale et la moyenne des sinistres. Cela peut être dû à des valeurs extrêmes qui impactent les valeurs estimés à la hausse. Nous décidons d'estimer à nouveau notre modèle en supprimant les valeurs influentes diagnostiquées grâce aux distances de Cook.

	Moyenne	Somme totale
Prédictions	18 518	299 430 991
Valeurs observées	16 969	274 391 537

Nous constatons que l'écart entre les valeurs prédites et observées s'est réduit, passant de 31% à 9.1 %

Nous avons mentionné le fait que des valeurs extrêmes pouvaient influer sur le modèle Gamma. Nous décidons de voir quel serait les résultats si nous décidions d'enlever les 4 % sinistres les plus élevés de notre base :

	Moyenne	Somme totale
Prédictions	17 122	276 865 967
Valeurs observées	16 969	274 391 537

L'écart n'est plus que 0.4 % entre la somme totale prédite et la somme totale observée. Il serait ainsi possible d'envisager un modèle séparé en deux parties, un modélisant les sinistres extrêmes, l'autre modélisant les sinistres d'un montant moins élevé.

Nous proposons de tester la sensibilité de la prédiction du modèle par rapport à une variation de l'échantillon de test et d'apprentissage. Nous simulons 500 fois le modèle avec des bases d'apprentissage et de validation différentes, nous retirons les valeurs influentes selon les distances de Cook et les 4 % sinistres les plus élevés. Nous calculons à chaque fois le rapport entre la somme prédite et la somme observée. Nous obtenons les résultats suivants :

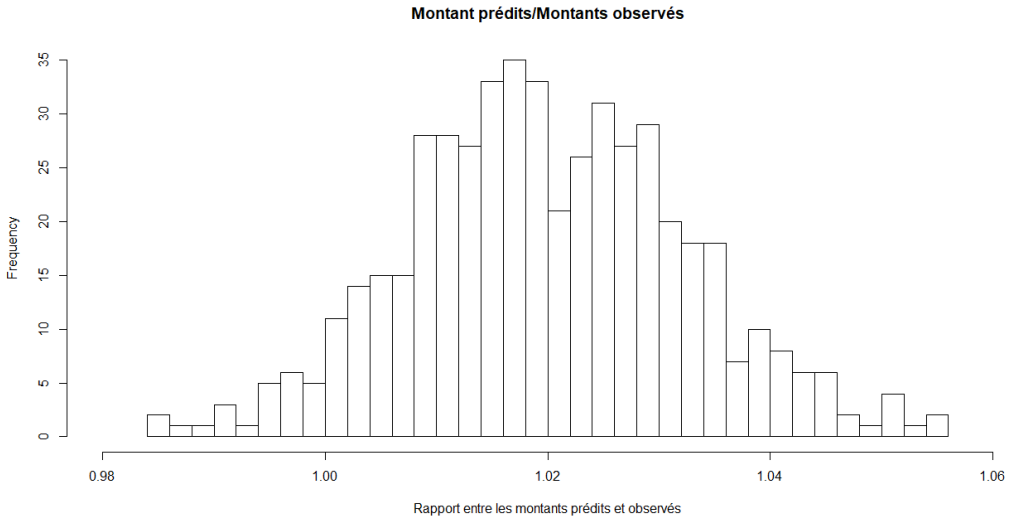


FIGURE 12: Rapport entre la somme prédite et la somme observée sur l'échantillon de validation, avec suppression des valeurs influentes et supérieures au quantile 0.96

Nous constatons que sur les 500 modélisations, l'écart entre la somme prédite et la somme observée est plutôt centré autour de 2 %. Nous décidons d'appliquer cette méthode à nouveau, en prenant le quantile à 95 %. Nous obtenons les résultats suivants :

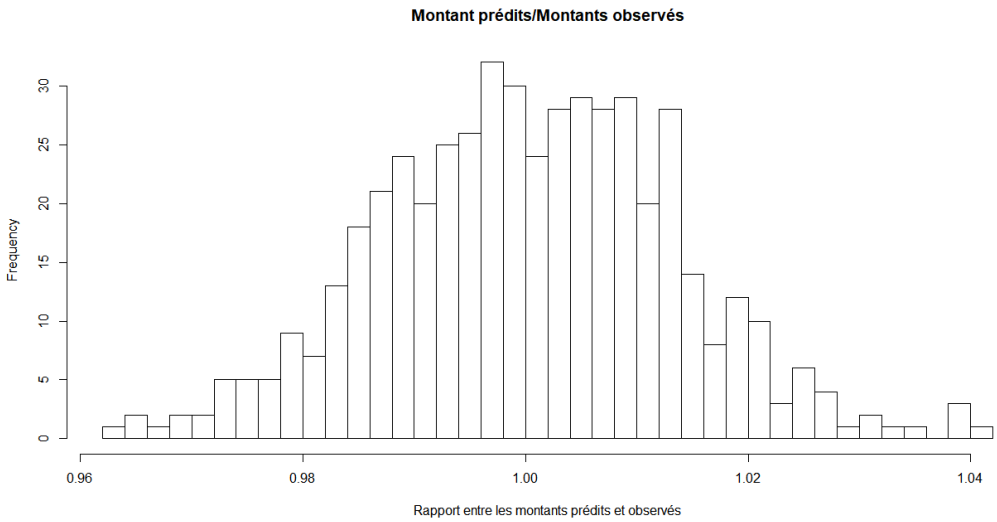


FIGURE 13: Rapport entre la somme prédite et la somme observée sur l'échantillon de validation, avec suppression des valeurs influentes et supérieures au quantile 0.95

Ici, l'écart entre la somme prédite et la somme observée est plutôt centré autour de 0%, ce qui peut nous amener à conclure que la partition effectuée grâce au quantile à 95 % est pertinente.

TABLE 6: Résultats des trois modèles Gamma testés

	Dependent variable :		
	Damage		
	(1)	(2)	(3)
CRS	-0.877^{***}	-2.783^{***}	-2.986^{***}
	(0.100)	(0.102)	(0.107)
ZoneInoAE	0.054^{***}	0.189^{***}	0.180^{***}
	(0.010)	(0.010)	(0.011)
ZoneInoX	-0.452^{***}	-0.607^{***}	-0.671^{***}
	(0.012)	(0.012)	(0.012)
EtageDeux étages	-0.361^{***}	-0.467^{***}	-0.509^{***}
	(0.009)	(0.009)	(0.009)
EtageTrois étages	-0.421^{***}	-0.811^{***}	-0.817^{***}
	(0.012)	(0.012)	(0.012)
EtageSplit-level	-0.755^{***}	-1.323^{***}	-1.270^{***}
	(0.021)	(0.021)	(0.021)
EtageMobil-home	0.277^{***}	0.196^{***}	0.237^{***}
	(0.024)	(0.024)	(0.024)
ResidencePrimY	-0.221^{***}	-0.195^{***}	-0.185^{***}
	(0.013)	(0.013)	(0.014)
Constant	-0.762^{***}	-0.864^{***}	-0.909^{***}
	(0.015)	(0.015)	(0.016)
Note :			*p<0.1 ; **p<0.05 ; ***p<0.01

Modèle avec une distribution lognormale

Nous nous intéressons au modèle Log-normal afin de voir si l'on peut obtenir un modèle valable.

Soit Y une variable suivant une loi log-normale. Sa densité peut s'écrire :

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} exp\left(-\frac{(lny-\mu)^2}{2\sigma^2}\right) \forall y \in \mathbb{R}^+$$

Cette densité ne peut s'écrire sous la forme d'une densité appartenant à une famille exponentielle. On peut néanmoins utiliser le fait que si Y suit une loi log-normale alors $log(Y)$ suit une loi normale, de ce fait nous retombons sur une loi appartenant à la famille exponentielle. Ainsi on pose $Y = exp(Y*)$ avec $Y \sim \mathcal{N}(\mu, \sigma^2)$. On remarquera que $E(Y) = E(exp(Y*)) \leq exp(E(Y*)) = exp(\mu)$

On a $E(Y) = exp(\mu + \sigma^2/2) = exp(\sigma^2/2)exp(E(Y*))$

En considérant toujours la variable "Valeur Assurée" comme une variable offset afin d'étudier le taux de destruction nous obtenons le modèle suivant :

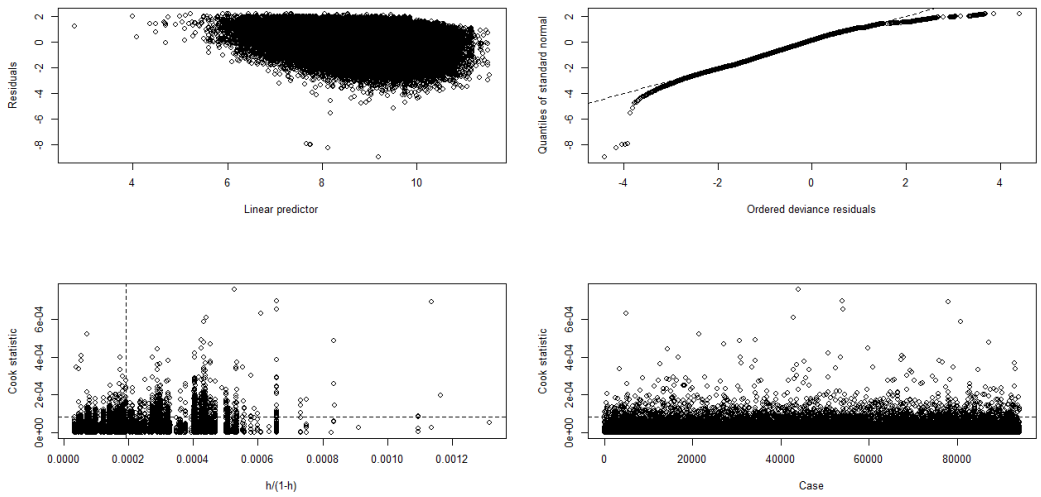


FIGURE 14: Graphiques de diagnostic du modèle log-normal

Nous remarquons toujours une structure sur le graphique des résidus par rapport aux valeurs ajustées. Les résidus de déviance sont alignés sur la droite d'Henry exceptée sur les valeurs extrêmes. Nous constatons toujours des valeurs influentes sur le modèle grâce aux distances de Cook.

	Moyenne	Somme totale
Prédictions	30 750	497 221 046
Valeurs observées	16 969	274 391 537

La somme des valeurs prédites est 81 % plus élevée que les valeurs observées.

Nous décidons de retirer les variables influentes du modèle, dans le même principe que ce que nous avons fait sur la régression Gamma. Nous obtenons :

	Moyenne	Somme totale
Prédictions	28 910	467 472 071
Valeurs observées	16 701	270 360 949

Nous constatons toujours une surestimation des valeurs observées de l'ordre de 70 %.

Les valeurs prédites restent très élevées par rapport aux observations. Nous avons fait l'hypothèse que la valeur assurée était une variable offset, à savoir dire que son coefficient est fixé à 1. C'est une hypothèse forte qui peut être remise en question. Nous choisissons de relancer nos modèles en retirant cette hypothèse sur le coefficient.

Modèle Gamma

	Dependent variable :
	Damage
CRS	0.450^{***}
	(0.104)
ZoneInoAE	0.065^{***}
	(0.011)
ZoneInoX	-0.203^{***}
	(0.013)
EtageDeux étages	-0.262^{***}
	(0.009)
EtageTrois étages	-0.329^{***}
	(0.012)
EtageSplit-level	-0.631^{***}
	(0.021)
EtageMobil-home	0.008
	(0.025)
ResidencePrimY	-0.295^{***}
	(0.014)
log((val))	0.628^{***}
	(0.004)
Constant	3.182^{***}
	(0.050)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Table 7: Résultat du modèle Gamma sans variable offset

	Moyenne	Somme totale
Prédictions	17 181	278 127 661
Valeurs observées	16 701	270 360 949

Table 8: Prédiction du modèle Gamma sans variable offset

L'écart entre la somme des valeurs prédites et des valeurs observées est de 2.8 % (Table 8), il était de 31% sur notre premier modèle Gamma. Nous constatons que le coefficient estimé pour la variable Valeur Assurée, une fois levée la contrainte de l'offset, est de 0.63. On peut ainsi penser que la contrainte du coefficient égal à 1 pouvait affecter le modèle. Le même exercice sur la régression Lognormale, appliquée à notre échantillon de validation nous donne les résultats suivants :

	Moyenne	Somme totale
Prédictions	20 568	332 958 136
Valeurs observées	16 701	270 360 949

Table 9: Prédiction du modèle Lognormal sans variable offset

On constate une surestimation de la somme totale des sinistres observés de 23 %. Nous rappelons que l'écart était de 81 % dans notre modèle Lognormale avec une variable "Valeur assurée" en offset.

Nous décidons de continuer avec le modèle Gamma. Nous déterminons le rapport entre sa déviance standardisée et le degré de liberté. Dans l'idéal se rapport doit être proche de 1, il est dans notre cas à 1.14.

Déviance résiduelle	155 619
Paramètre de dispersion	1.44
Déviance standardisée	108 068
Degré de liberté	94071
Rapport Déviance standardisée/Degré de liberté	1.14

Déviance standardisée du modèle Gamma

Étude des variables explicatives

La variable CRS a un coefficient positif ce qui peut sembler contre-intuitif. Lorsque nous avions mis la variable "Valeur assurée", ce coefficient était négatif. Ce qui peut s'interpréter par le fait que les mesures de prévention réduisent le coût d'un sinistre lorsque celui-ci se produit. Dans notre dernier modèle, le coefficient est positif. Une interprétation possible est que les communes ayant un CRS élevé, sont celles qui sont le plus exposées aux risques, d'où le besoin d'adopter des mesures de prévention. Le manque de variables explicatives ou une mauvaise spécification du modèle peut être également la cause d'un coefficient positif sur cette variable.

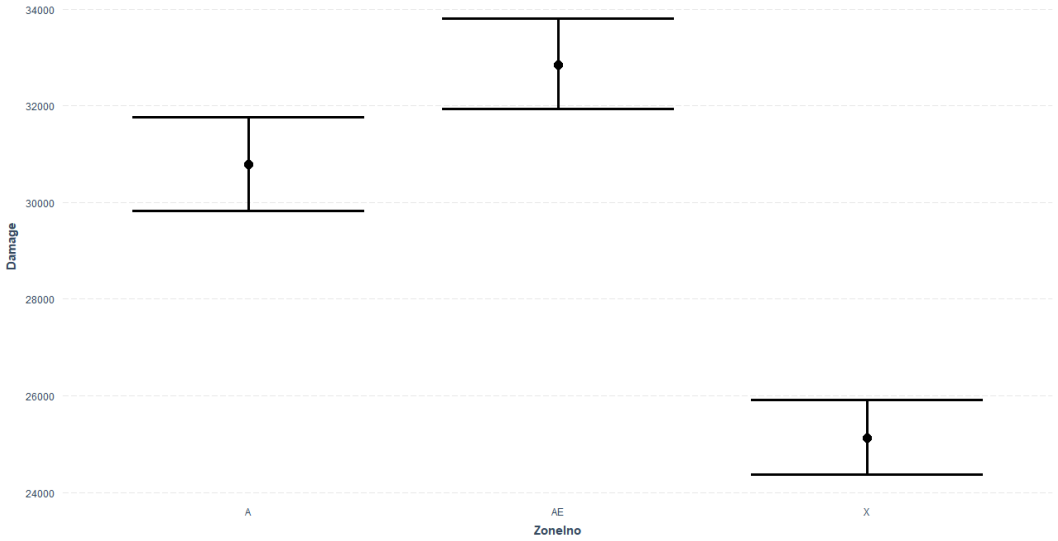


FIGURE 15: Montant moyen des sinistres attendus en fonction du type de zone à risque

Les autres coefficients ont une interprétation plus simple. La variable Zone inondable nous fait apparaître clairement la différence de moyenne des sinistres attendus lorsqu'une habitation se trouve en zone inondable ou non. Nous constatons qu'il y a un léger écart entre la moyenne attendue pour zones A et AE (6.7%). L'écart est de 19.4 % entre la moyenne ajustée d'un sinistre en zone à fort risque sans BFE (Zone A) par rapport à un sinistre en zone à faible risque.

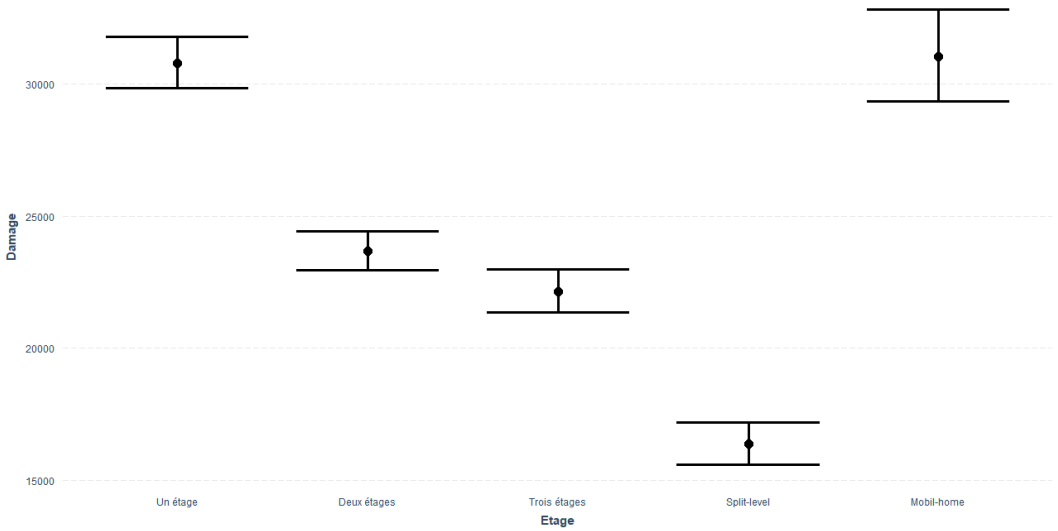


FIGURE 16: Montant moyen des sinistres attendus en fonction du nombre d'étages

Nous voyons également que le nombre d'étages de la maison a un impact sur la moyenne des sinistres attendus. Le montant est plus élevé pour les maisons d'un

étage. On peut interpréter cela par le fait que le taux de destruction est plus élevé pour les maisons d'un étage, le contenu assuré ne pouvant être à l'abri dans les étages supérieurs. La catégorie Mobilhome n'est pas significative, on peut considérer dans notre modèle que la catégorie "Un étage" et "Mobil-home" comme similaire.

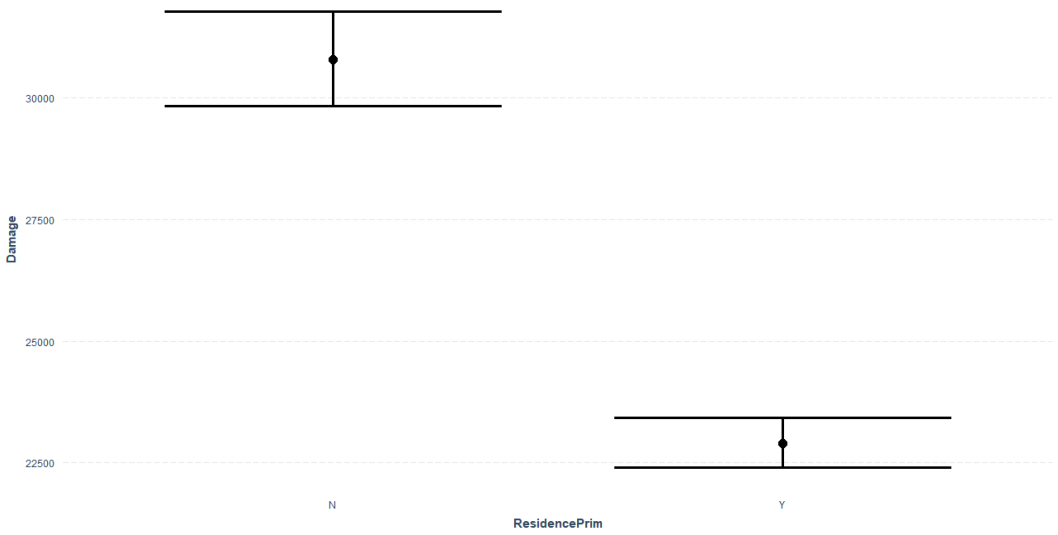


FIGURE 17: Montant moyen des sinistres attendus pour les résidences principales et secondaires

On observe également une différence entre le montant des sinistres attendus pour les résidences principales et secondaires. Le montant est 25 % plus faible pour les résidences principales. On peut interpréter cela par le fait que les propriétaires de résidences principales investissent plus dans la prévention du risque que pour leur résidences secondaires.

Comparaison avec les sinistres d'un événement extrême : l'ouragan Katrina

Nous sélectionnons les sinistres qui ont été déclarés entre le 28 et 31 août 2005 en Floride et Louisiane. Nous avons une base de 135 000 sinistres. Nous estimons à nouveau notre modèle Gamma avec cette nouvelle base. Nous obtenons les résultats suivants :

TABLE 10: Résultat de la régression Gamma des deux sources de données

	Dependent variable :	
	Damage	
	Ouragan Katrina	Historique des États du bassin du Mississippi
CRS	-3.821^{***}	0.450^{***}
	(0.040)	(0.104)
ZoneInoAE	-0.018^{*}	0.065^{***}
	(0.009)	(0.011)
ZoneInoX	-0.484^{***}	-0.203^{***}
	(0.010)	(0.013)
ZoneInoV	-0.390	
	(0.322)	
ZoneInoVE	-0.291^{***}	
	(0.018)	
EtageDeux étages	-0.080^{***}	-0.262^{***}
	(0.004)	(0.009)
EtageTrois étages	-0.039^{***}	-0.329^{***}
	(0.010)	(0.012)
EtageSplit-level	-0.032	-0.631^{***}
	(0.021)	(0.021)
EtageMobil-home	-0.350^{***}	0.008
	(0.025)	(0.025)
ResidencePrimY	-0.032^{***}	-0.295^{***}
	(0.004)	(0.014)
log((Valeur Assurée))	0.793^{***}	0.628^{***}
	(0.003)	(0.004)
Constant	2.506^{***}	3.182^{***}
	(0.031)	(0.050)
Note :		*p<0.1 ; **p<0.05 ; ***p<0.01

Déviance résiduelle	104 279
Paramètre de dispersion	0.31
Déviance standardisée	336 383
Degré de liberté	115 420
Rapport Déviance standardisée/Degré de liberté	2.91

Déviance standardisée du modèle Gamma sur les données de l'ouragan Katrina

En comparant le modèles issus des sinistres de l'ouragan Katrina et ceux du bassin du Mississippi sur tout l'historique, nous constatons qu'il y a un moins d'écart entre les coefficients des variables explicatives : la sinistralité attendue est moins dépendante du nombre d'étage de la maison ou du fait que celle-ci soit une résidence principale ou secondaire. Nous remarquons que le coefficient de la variable CRS est négatif, ce qui signifie que la sinistralité modélisée par la régression est moins élevée dans les communes avec un haut niveau de mesures de prévention.

Néanmoins, en analysant les données de la régression, nous constatons que la dispersion du modèle peut remettre en cause celui-ci. Alors que nous souhaitons un rapport entre déviance standardisée et degré de liberté proche de 1, nous obtenons un rapport à 2.91. La forte dispersion des valeurs peut expliquer cet effet.

Variance des sinistres de l'ouragan Katrina	5 445 909 201
Variance des sinistres des États du bassin du Mississippi	806 899 661

4 Chapitre IV : Analyse du taux de destruction

4.1 Analyse du portefeuille

Sachant que la prime est calculée en appliquant un pourcentage de la valeur assurée, il est intéressant de modéliser le taux de destruction. Nous avons vu que les régressions Lognormale et Gamma étaient moins précises pour estimer la moyenne des sinistres observés lorsqu'il y a la contrainte de l'offset "Variable assurée" (pour retomber sur un taux de destruction). Nous allons étudier ici la régression Bêta pour modéliser ce taux.

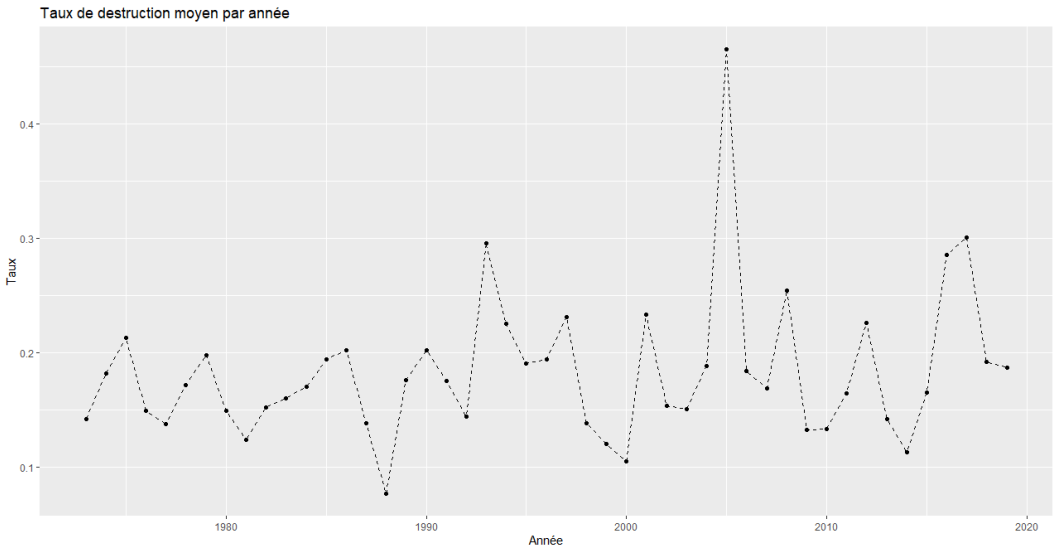


FIGURE 18: Taux de destruction moyen par année entre 1973 et 2019

Au vu de l'allure de la courbe du taux de destruction il est intéressant de comparer ce taux avec le nombre de sinistres par année.

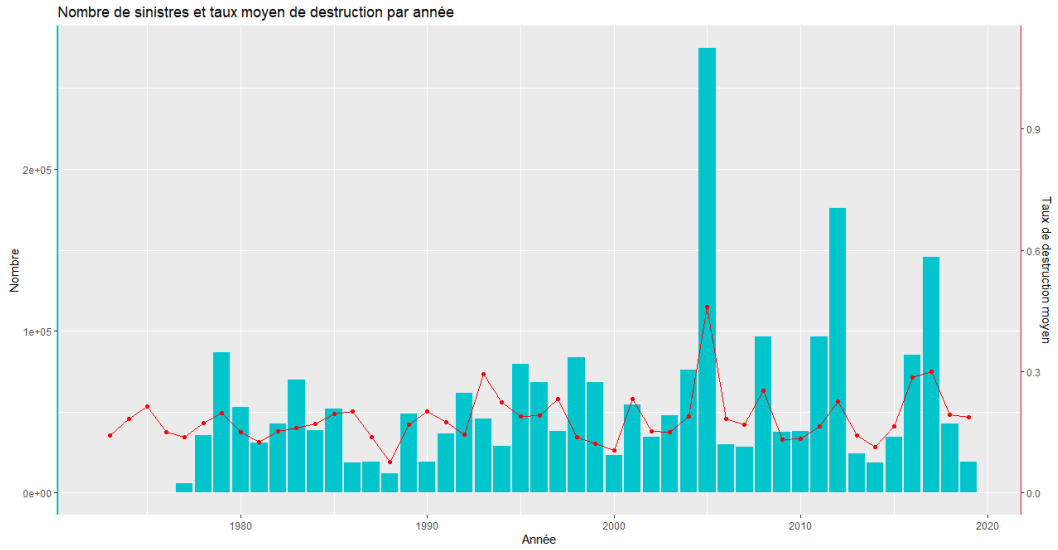


FIGURE 19: Taux de destruction moyen par année entre 1973 et 2019

Le taux de destruction à tendance à augmenter lors des années présentant un grand nombre de sinistres et à diminuer lorsqu'il y a moins de sinistres. On peut étudier la corrélation entre les deux variables. En calculant le coefficient de Pearson on obtient une valeur $\rho = 0.7$.

Variable "Etage"

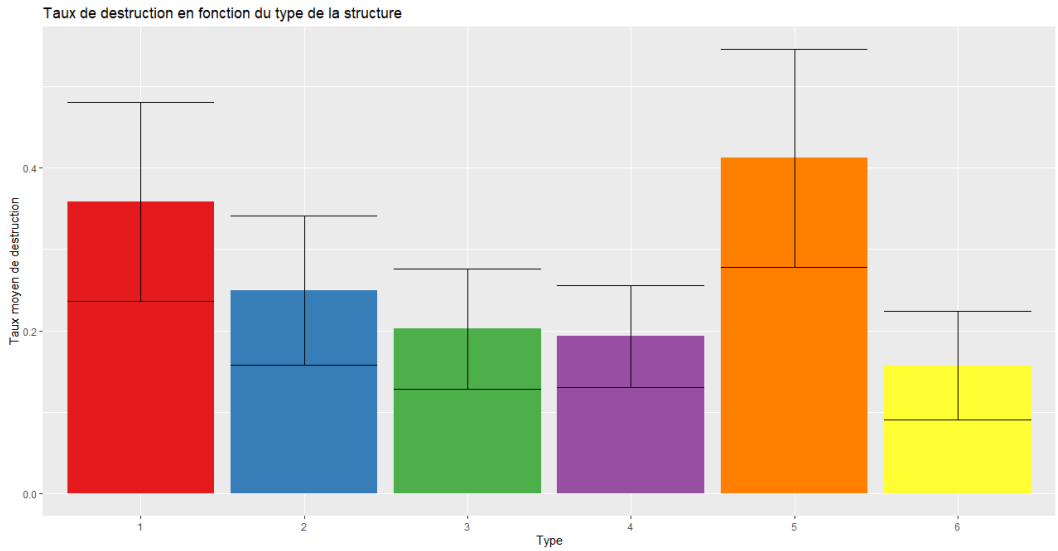


FIGURE 20: Taux de destruction moyen par type d'étages

Nous rappelons que les types d'étages sont répertoriés en 6 catégories :

— 1= Maison à un étage
— 2= Maison à deux étages
— 3= Maison à trois étages ou plus
— 4= Maison "split-level"
— 5= Mobilhome
— 6= Maison de ville avec trois étages au moins

On observe la fort taux de destruction moyen des mobilhomes et des maisons à un étage beaucoup plus exposés aux inondations que les structures plus hautes.

Variable "Zone inondable"

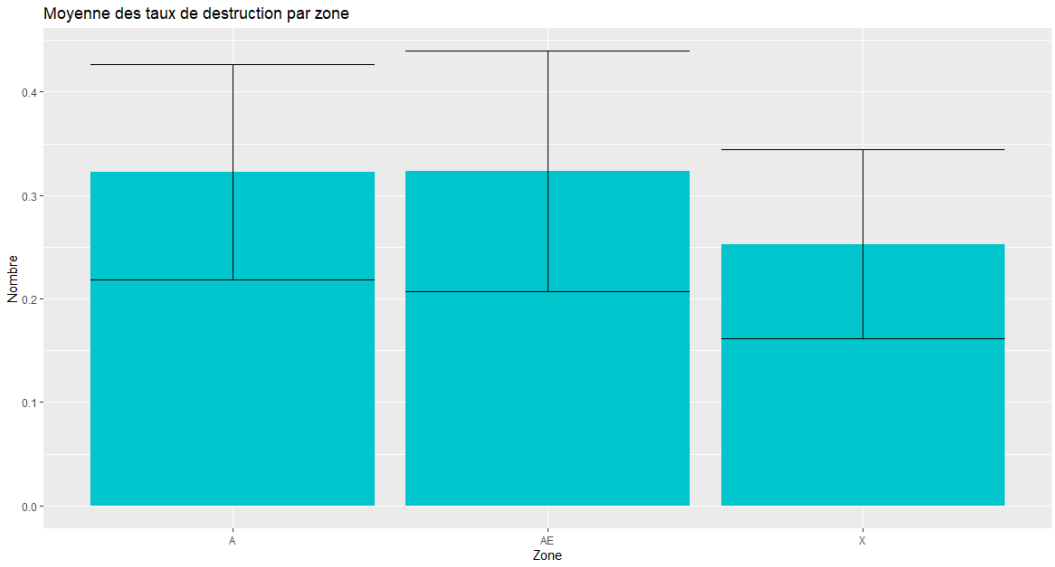


FIGURE 21: Taux de destruction moyen par zone inondable

En regroupant les zones A1-A30 au sein de la zone AE pour une lecture simplifiée, on observe une différence notable dans les taux de destructions entre les maisons se situant en zone à risque faible et celles se situant dans les zones à fort risque.

Variable "Sous-Sol"

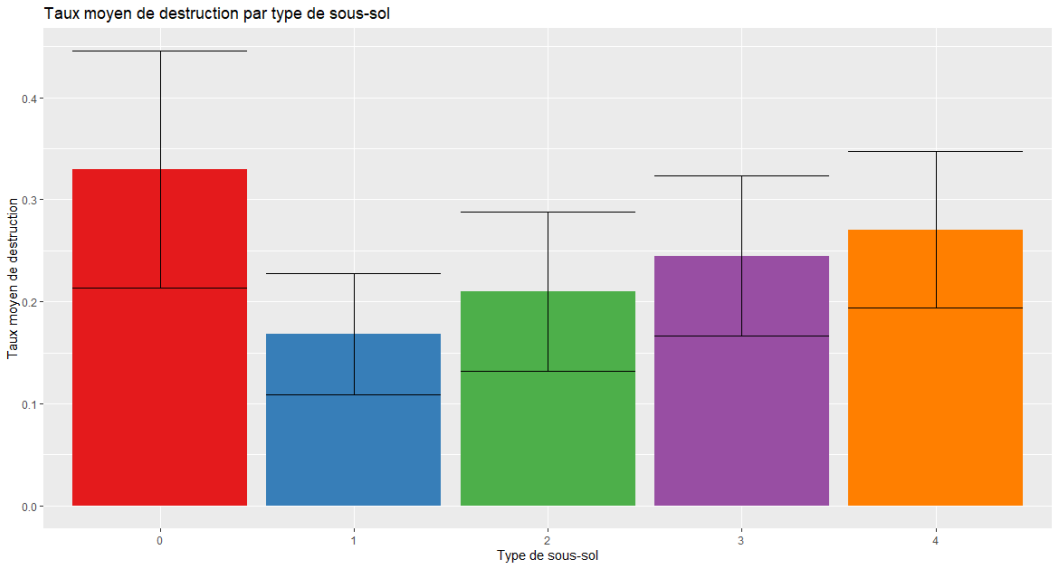


FIGURE 22: Taux de destruction moyen par type de sous-sol

Nous rappelons que les codes attribués aux types de sous-sol sont les suivants : 0= Absence de sous-sol ; 1 = Sous-sol fini ; 2 = Sous-sol non-fini ; 3= Vide sanitaire ; 4 =Vide sanitaire en sous-sol. Les maisons sans sous-sol connaissent un taux de destruction plus élevées que les autres.

Variable "Année de construction"

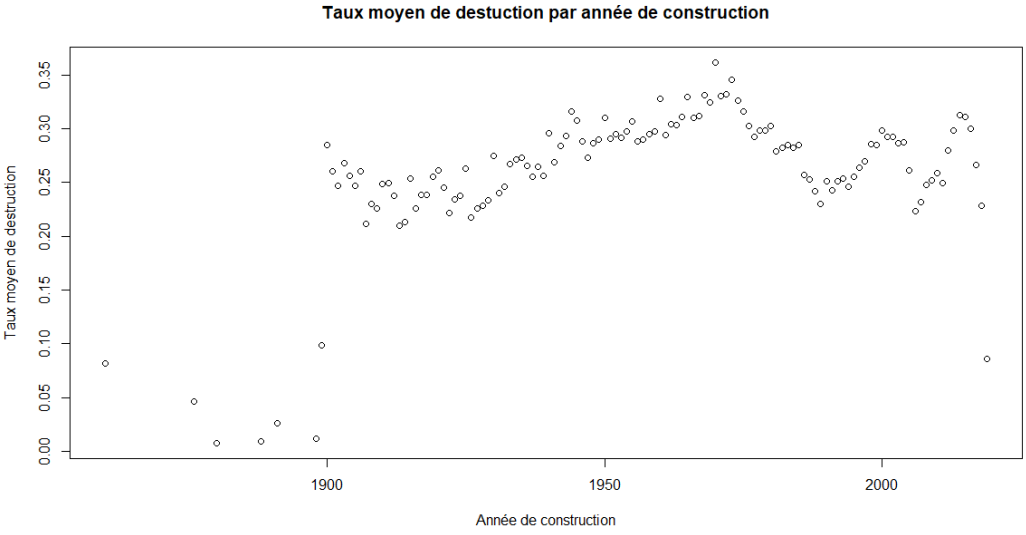


FIGURE 23: Taux de destruction moyen par année de construction

Il ne semble pas y avoir de tendance visible entre l'année de construction et le taux de destruction.

Variable "Résidence primaire"

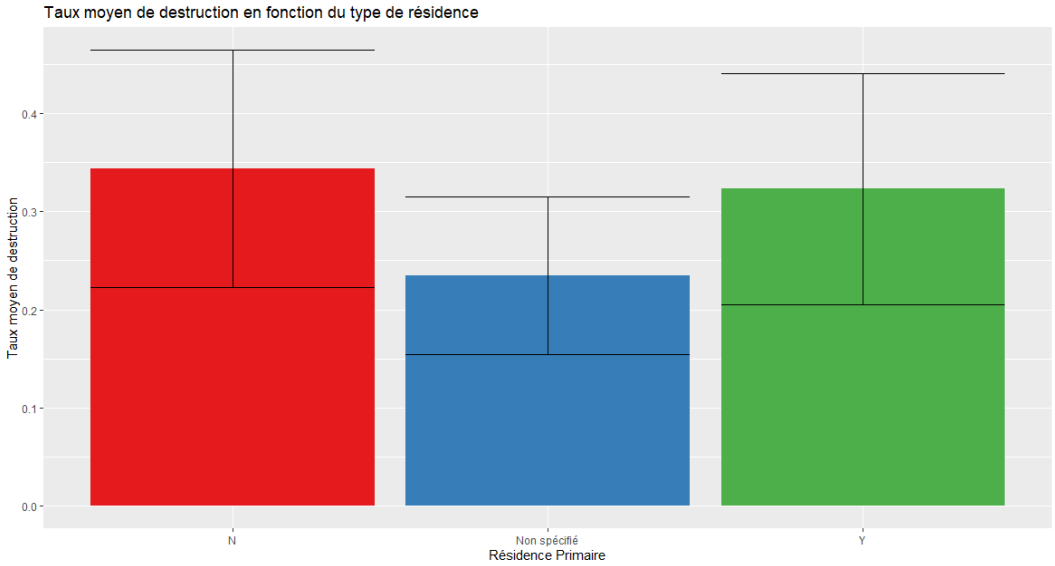


FIGURE 24: Taux de destruction moyen par type de résidence

On peut émettre l'hypothèse que les sinistres dont la résidence principales n'a pas été spécifiée sont des résidences principales. On remarquerait donc que les maisons secondaires possèdent un taux de destruction plus élevée que les résidences primaires. L'absence de moyens de prévention aussi importants que sur les résidences principales peut expliquer ce phénomène.

4.2 Regression bêta

4.2.1 Théorie

Lorsque l'on cherche à faire une régression avec une variable réponse comprise entre 0 et 1, on est souvent amené à faire une transformation de la variable afin d'avoir un intervalle de régression dans $\mathcal{R}$. Une des transformations les plus utilisées est la composition par la fonction $logit$: $f(x) = ln(\frac{y}{1-y})$ Un des problèmes posé par ce type d'approche est que l'on interprète les paramètres de régression selon $f(y)$ et non pas selon la variable originale y (Cribari-Neto et Zeileis). Une solution proposée par Silvia L.P Ferrari et Francisco Cribari-Neto (2004) est d'estimer une régression pour une variable distribuée selon une loi bêta. La loi bête possède l'avantage d'être particulièrement flexible :

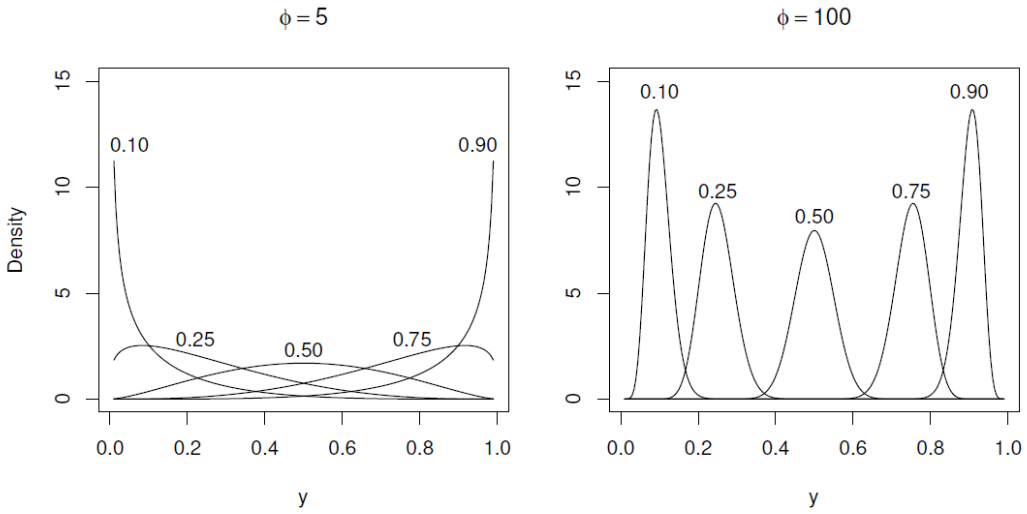


FIGURE 25: Distributions de lois bêta avec $\mu = 0.1, 0.25, 0.5, 0.75$ et 0.9 et $\phi = 5$ et 100. Source : Cribari-Neto et Zeilis (2010)

La densité de la loi bêta s'écrit :

$$f(y, p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}$$

avec $0 < y < 1, p > 0, q > 0$ et $\Gamma(x)$ la fonction gamma.

$$E(y) = \frac{p}{(p+q)}$$

et

$$V(y) = \frac{pq}{(p+q)^2(p+q+1)}$$

La paramétrisation proposée par Ferrari et Cribari-Neto est différente : on cherche à modéliser la moyenne de la variable réponse. On pose donc $\mu = \frac{p}{p+q}$ et $\phi = p+q$. Ce qui nous donne :

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi))} y^{\phi\mu-1}(1-y)^{(1-\mu)\phi-1}$$

pour $0 < y < 1$, avec $0 < \mu < 1$ et $\phi > 0$ On note $y \quad \mathcal{B}(\mu,\phi)$ avec $E(y) = \mu$ et $V(y) = V(\mu)/(1+\phi) = \mu(1-\mu)/(1+\phi)$

Soit $y_1, ...y_n$ n valeurs observées de la variable réponse de moyenne μ_i et d'un paramètre de dispersion ϕ

$$g(\mu_i) = x_i^T\beta = x_i \sum_{j=1}^{k} x_i j\beta_i = \eta_i$$

avec $\beta = (\beta_1, ..., \beta_k)^T$ un vecteur de paramètres à estimer et $x_{t1}, ..., x_{tk}$ les k valeurs des variables explicatives pour les n observations. La fonction lien est définie par $g(x)$. Elle est strictement monotone et deux fois dérivables, est dont l'image de l'intervalle $]0,1[$ est $\mathcal{R}$ La variance de y étant une fonction de μ, on obtient un modèle hétéroscedastique avec :

$$V(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi} = \frac{g^{-1}(x_i^T\beta)[1-g^{-1}(x_i^T\beta)]}{1+\phi}$$

La fonction de log-vraisemblance $l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi)$ avec :

$l_i(\mu_i,\phi) = ln(\Gamma(\phi)) - log(\Gamma(\mu_i\phi)) - log(\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)ln(y_i) + [(1-\mu_i)\sigma - 1]ln(1-y_i)$

Pour étudier l'ajustement du modèle, on utilise les résidus de Pearson, nommé *standardized ordinary residuals* par les auteurs :

$$\frac{y_i - \hat{\mu}_i}{\sqrt{V\hat{A}R(y_i)}}$$

avec $V\hat{A}R(y_i) = \hat{\mu_i}(1-\hat{\mu_i}/(1+\hat{\phi_i}), \hat{\mu_i} = g_1^{-1}(x_i^T\hat{\beta})$ et $\hat{\phi_i}(z_i^T\hat{\gamma})$.

4.2.2 Application

Nous cherchons à expliquer le taux de destruction en fonction des variables explicatives présentes dans notre base de données. Nous prenons les taux de destruction sur l'année 2008. Pour les besoins du modèle nous séparons les taux de destruction égaux à 1 et les taux strictement inférieurs à 1. Nous observons sur le graphique une masse de Dirac en 1, lorsque le sinistre atteint la valeur maximale assurée.

```
 Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.7016 -0.5722 -0.0033  0.5495  3.6025
```

```
Coefficients (mean model with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.76449    0.10851  -7.045 1.85e-12 ***
CRS               -0.28689    0.61353  -0.468  0.64006
ZoneInoAE          0.08636    0.10066   0.858  0.39094
ZoneInoX          -0.32302    0.10460  -3.088  0.00201 **
EtageDeux étages  -0.19466    0.08019  -2.428  0.01520 *
EtageTrois étages -0.16680    0.09368  -1.780  0.07500 .
EtageSplit-level  -0.40802    0.19814  -2.059  0.03947 *
EtageMobil-home   -0.10381    0.26575  -0.391  0.69608
ResidencePrimY    -0.28907    0.06786  -4.260 2.05e-05 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  2.45483    0.09568   25.66   <2e-16 ***
---
```

Nous remarquons que la variable CRS n'est pas significative. La Zone inondable AE n'est pas non plus significative, ce qui signifie qu'on peut la regrouper avec la zone inondable A. On remarque la baisse du montant des sinistres attendus sur les zones à faible risque (X) par rapport à la modalité de référence qui est la zone inondable A. Nous remarquons également que pour la variable étage, tous les coefficients sont négatifs, ce qui signifie que le taux de destruction attendu est plus élevé pour les maisons à un étage (la modalité de référence) que pour celles avec plusieurs étages. La modalité Mobilhome est non significative. Pour la variable "Résidence principale", comme pour les modèles Gamma et Lognormal, nous remarquons que le taux de destruction attendus est plus élevé pour les résidences secondaires.

L'étude du half-normal plot des résidus de déviance nous montre que ceux-ci ne sont pas dans l'intervalle de confiance des quantiles du distribution normale entre 1.5 et 2.5.

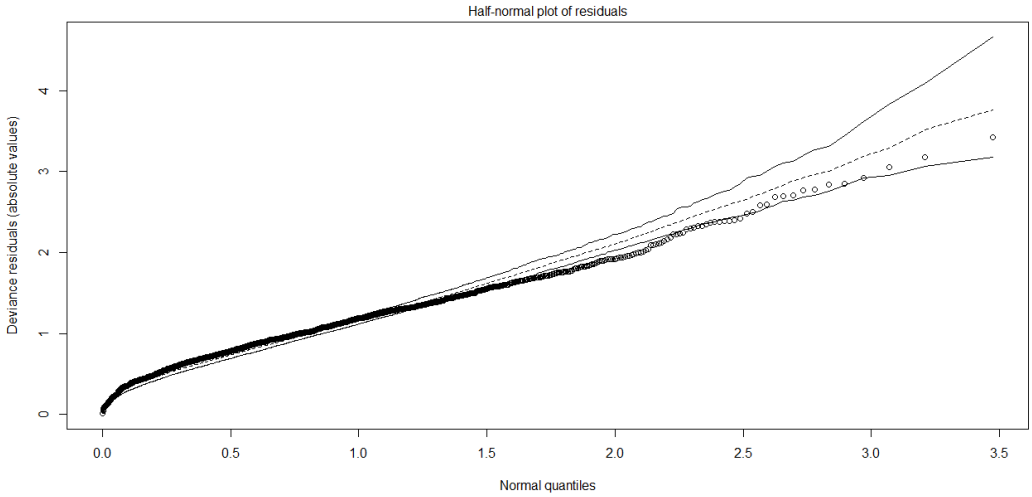


FIGURE 26: Half-normal plot des résidus de déviance d'un modèle Bêta avec un paramètre de dispersion unique.

Afin de prendre en compte un éventuel phénomène d'hétéroscédasticité, on peut ajouter des paramètres de régression supplémentaires à notre variable de dispersion ϕ. En calculant les paramètres de dispersion pour les variables de régression on obtient :

```
Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.3938 -0.5750 -0.0063  0.5669  3.4891

Coefficients (mean model with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.05714    0.08817 -11.990  < 2e-16 ***
CRS               -0.15881    0.51598  -0.308  0.75825
ZoneInoAE          0.08253    0.08478   0.973  0.33034
ZoneInoX          -0.39224    0.09212  -4.258 2.06e-05 ***
EtageDeux étages  -0.20727    0.06702  -3.093  0.00198 **
EtageTrois étages -0.17287    0.08024  -2.154  0.03121 *
EtageSplit-level  -0.75710    0.18644  -4.061 4.89e-05 ***
EtageMobil-home   -0.39189    0.23957  -1.636  0.10188
ResidencePrimY    -0.25677    0.05725  -4.485 7.28e-06 ***

Phi coefficients (precision model with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.60465    0.13198   4.581 4.62e-06 ***
CRS               -0.03069    0.76026  -0.040 0.967795
ZoneInoAE         -0.04381    0.12512  -0.350 0.726230
```

```
ZoneInoX              0.32417    0.13339   2.430 0.015089 *
EtageDeux étages      0.19223    0.09921   1.938 0.052668 .
EtageTrois étages     0.14665    0.11734   1.250 0.211404
EtageSplit-level      1.01086    0.27699   3.649 0.000263 ***
EtageMobil-home       0.58341    0.36244   1.610 0.107474
ResidencePrimY        0.13921    0.08475   1.643 0.100449
```

Résultat du modèle Bêta avec plusieurs paramètres de dispersion.

On observe que les paramètres de dispersion de l'intercept, de la zone X, des maisons à deux étages et split-level sont significatifs. Nous observons l'impact de l'ajout de ces paramètres de dispersion sur l'échantillon de validation :

	Moyenne	Variance
Echantillon de validation	22.5 %	6.3%
Prédiction(modèle Bêta avec un paramètre de dispersion unique)	23.1 %	0.25 %
Prédiction (modèle Bêta avec un paramètre de dispersion pour chaque variable)	22.6 %	0.5%

Nous voyons que le modèle a du mal à capter la variance de l'échantillon de validation. La moyenne des taux de destruction prédits est relativement proche de celui des taux observés. L'ajout des paramètres de dispersion améliore la précision de la moyenne prédite et double la variance des données (de 0.25 % à 0.5 %). Néanmoins, la variance reste très différente de celle de l'échantillon observée. Nous pouvons faire un test de rapport de vraisemblance pour voir dans quelle mesure l'ajout des paramètres de dispersion améliore notre modèle.

```
Likelihood ratio test

Model 1: TauxDest ~ CRS + ZoneIno + Etage + ResidencePrim
Model 2: TauxDest ~ CRS + ZoneIno + Etage + ResidencePrim
| ZoneIno + Etage
  #Df LogLik Df  Chisq Pr(>Chisq)
1  10 736.21
2  16 753.64  6 34.845  4.618e-06 ***
```

Résultat du test du rapport de vraisemblance

On observe que le modèle avec un paramètre de dispersion propre pour la variable "Zone inondable" et "Etage" est significativement meilleur que le premier. De plus, l'étude des résidus de déviance nous montre un meilleur ajustement des résidus par rapport à l'intervalle de confiance des quantiles d'une loi normale :

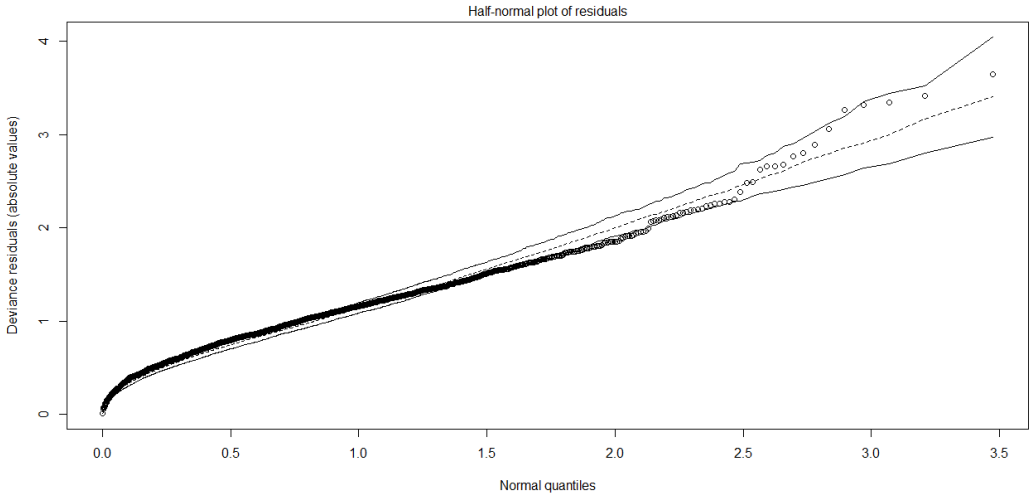


FIGURE 27: Half-normal plot des résidus de déviance d'un modèle Bêta avec plusieurs paramètres de dispersion.

4.3 Conclusion

Nous avons étudié la distribution des montants des sinistres sur les États du bassin du Mississippi. L'analyse des résultats sur le taux de destruction (i.e avec la variable "Valeur assurée" en offset) nous a montré que l'ajustement des modèles était très faible et nous faisais largement surestimer le risque. Les régressions linéaires faites en enlevant l'attribut offset à la variable "Valeur assurée" nous a permis d'avoir un modèle qui ajuste mieux la somme totale observée de notre échantillon de validation. Nous avons cependant dû supprimer des observations influentes et exclure les 5 % sinistres les plus élevés de notre base.

Afin de modéliser le taux de destruction, nous avons essayé d'ajuster un modèle avec une régression Bêta. Nous constatons que la dispersion des valeurs observées est mal prise en compte lorsque l'on ajuste notre modèle avec notre échantillon de validation.

5 Chapitre V : Autocorrélation spatiale

Dans cette partie nous nous intéresserons à la modélisation spatiale de la sinistralité, plus particulièrement sur l'autocorrélation spatiale de la sévérité. Nous avons vu que cette autocorrélation spatiale était une des causes du retrait des assureurs privés au début du XXè du marché de l'assurance inondation, nous cherchons donc à quantifier celle-ci.

5.1 Présentation théorique

Définition

L'autocorrélation spatiale est la corrélation d'une variable avec elle-même, à partir de ses observations présentes à différentes localisations géographiques. La présence d'une autocorrélation spatiale remet en cause l'hypothèse d'indépendance sur laquelle se fondent de nombreuses analyses, notamment l'hypothèse d'indépendance entre les sinistres X_i, dans un modèle collectif du type : $\sum_{i=1}^{n} X_i$. L'autocorrélation spatiale peut être positive lorsque les observations sont relativement proches de leur observations voisines ou négative lorsque les observations de la variable sont relativement opposées par rapport à leurs observations voisines. En absence d'autocorrélation spatiale, on peut admettre que la répartition géographique des valeurs de la variable est aléatoire.

Les indices d'autocorrélation spatiale nous permettent de quantifier la dépendance géographique d'une variable et de tester la significativité de celle-ci. Pour les déterminer, nous avons besoin de définir la notion de voisinage et la similitude des valeurs.

Deux notions sont à considérer lorsque l'on parle d'autocorrélation spatiale :

- L'autocorrélation spatiale globale dans un espace géographique donné.
- L'autocorrélation spatiale locale qui mesure la corrélation locale d'une observation par rapport à ses observations voisines.

Diagramme de Moran

Le diagramme de Moran permet de comprendre rapidement la structure spatiale du phénomène observé. On détermine une matrice de poids normalisée W qui permet de définir les observations voisines. Le diagramme représente un nuage de point avec les valeurs de la variable observée centrée en abscisse et les valeurs moyennes des observations voisines $W * y$ en ordonnée. La moyenne empirique de $W * y$ est égale à celle de y donc à 0. Afin de délimiter des quadrants on trace les droites d'équations $y = 0$ et $W * y = 0$. On détermine également la droite de régression linéaire de $W * y$ en fonction de y. S'il n'y a pas de relation entre y et $W * y$ alors la pente de régression est nulle. Sinon la pente de régression est non nulle et il y a

une corrélation entre y et Wy. Les quadrants délimitent des types de dépendance spatiales particulières :

- Les observations en haut à droite, représentent les valeurs de la variable plus élevées que la moyenne et dont les observations voisines ont des valeurs proches.
- Les observations en bas à gauche, représentent les valeurs de la variable moins élevées que la moyenne et dont les observations voisines ont des valeurs proches.
- Les observations en bas à droite, représentent les valeurs de la variable plus élevées que la moyenne et dont les observations voisines ne sont pas proches.
- Les observations en haut à gauche, représentent les valeurs de la variable moins élevées que la moyenne et dont les observations voisines ne sont pas proches.

Exemple :

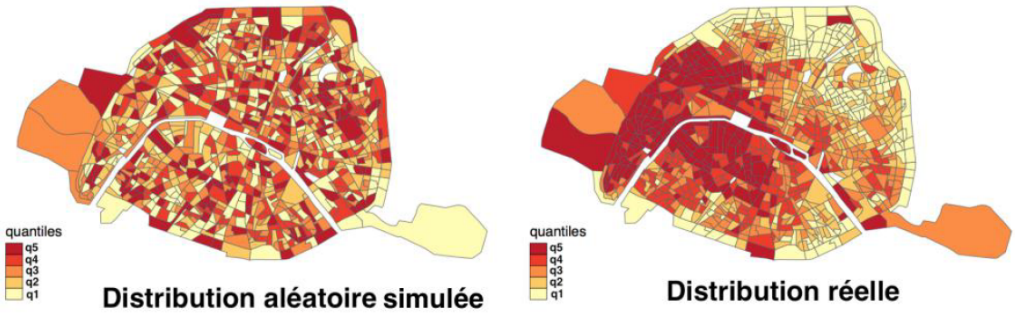


FIGURE 3.1 – Illustration, sur les Iris parisiens, de l'écart entre une distribution aléatoire et une distribution autocorrélée spatialement
Source : *Insee, Revenus Fiscaux Localisés 2010*

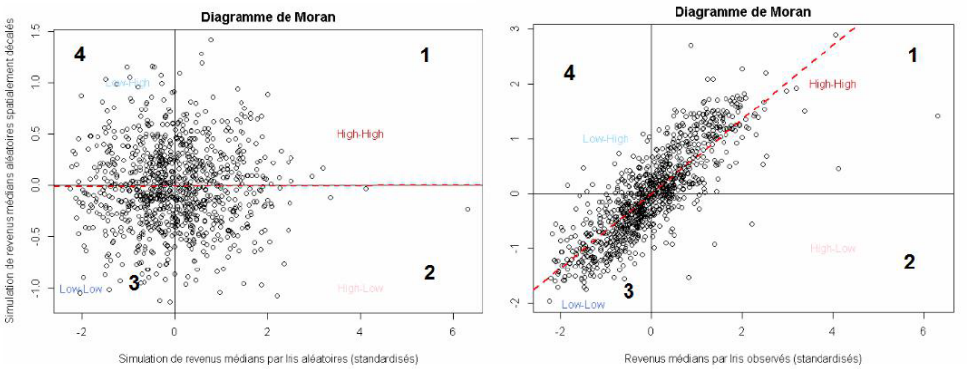


FIGURE 3.2 – Diagramme de Moran des revenus médians par Iris standardisés et d'une simulation de répartition aléatoire des revenus médians par Iris, pour les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

FIGURE 28: Diagramme de Moran sur les revenus médians parisiens. Insee 2018

5.2 Indices d'autocorrélation spatiale

Les indices d'auto-corrélation spatiale nous permettent de savoir si l'observation des valeurs dans l'espace n'est pas aléatoire et de quantifier cette auto-corrélation ainsi que sa significativité. On cherche à tester l'hypothèse suivante :

- H_0 : Absence d'auto-corrélation spatiale
- H_1 : Présence d'auto-corrélation spatiale

Deux hypothèses sont possibles pour la distribution de la variable observée y sous H_0

- **Hypothèse de normalité** : Les variables y_i suivent une loi normale propre à chaque unité géographique i.
- **Hypothèse de randomisation** : On compare la statistique obtenue avec la distribution de la statistique déterminée à partir de la permutation aléatoire des données. S'il n'y a pas d'autocorrélation spatiale, alors les combinaisons possibles des observations sont équiprobables. En déterminant la distribution aléatoire de la statistique, on peut déterminer un intervalle à partir duquel on considère que l'autocorrélation spatiale est significative.

Les indices d'auto-corrélation spatiale se calculent généralement de la manière suivante :

$$Corr(Y, WY) = \frac{Cov(Y, WY)}{\sqrt{(Var(Y) * Var(WY)}}$$

où W est une matrice de pondération calculée de la manière suivante :

Soit D une matrice telle que $d_{ij} = 1$ si l'unité géographique i et voisine avec l'unité j et $d_{ij} = 0$ sinon. On note W tel que $w_{ij} = \frac{d_{ij}}{\sum_j d_{ij}}$

Deux indices sont utilisés pour tester l'autocorrélation spatiale : l'indice de Moran et l'indice de Geary. Le premier prend en compte les variances et covariances, le second prend en compte la différence entre les observations voisines. On considère que l'indice de Moran est plus stable que celui de Geary.

Indice de Moran

Soit I_W l'indice de Moran tel que :

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} avec i \neq j$$

Si I_W est proche de 0 alors il n'y a pas d'évidence d'autocorrélation spatiale. Si I_W est significativement supérieur ou inférieur à 0 alors on peut admettre l'existence d'une autocorrélation spatiale positive ou négative. Nous devons établir un seuil c_0 tel que si $|I| > c_0$ alors on rejette l'hypothèse nulle d'absence d'auto-corrélation. La démarche du test est la suivante :

— On calcule l'indice de Moran avec les valeurs observées I^1
— On effectue N permutations aléatoires des valeurs et pour chacune de ces permutations, on calcule l'indice de Moran I^j.
— Sous H_0, les permutations sont équiprobables. On calcule la p-value telle que :

$$p = \frac{Card\{I^{(j)} > I^{(1)}\}}{N+1}$$

On rejette l'hypothèse nulle si $p < \alpha$

5.2.1 Application

Nous prenons la moyenne des taux de sinistres par comté dans les États du bassin du Mississippi et nous cherchons à savoir s'il y a une autocorrélation spatiale entre les comtés.

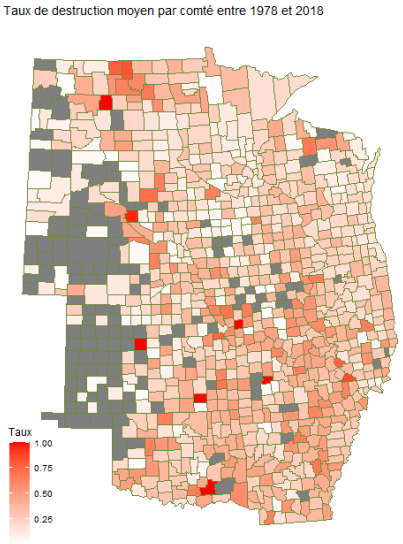


FIGURE 29: Taux de destruction moyen par comté entre 1973 et 2018

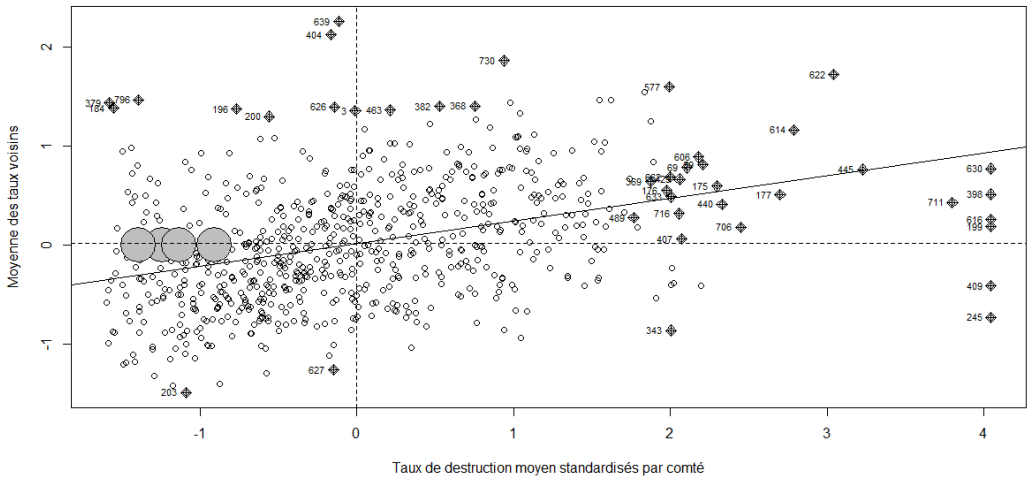


FIGURE 30: Diagramme de Moran des taux de destruction par comté

Nous déterminons d'abord la matrice de pondération W. Ensuite nous traçons le diagramme de Moran pour avoir une idée de la structure spatiale du phénomène observé.

On observe une auto-corrélation positive, avec une certaine concentration de valeurs dans le quadrant en bas à gauche, c'est-à-dire les valeurs moins élevées que la moyenne et dont les observations voisines ont des valeurs proches.

Nous cherchons à tester la significativité de cette autocorrélation. Nous simulons 1000 permutations des taux de destruction sur l'ensemble de l'espace donné et nous comparons la distributions observée avec l'indice de Moran sur les valeurs observées.

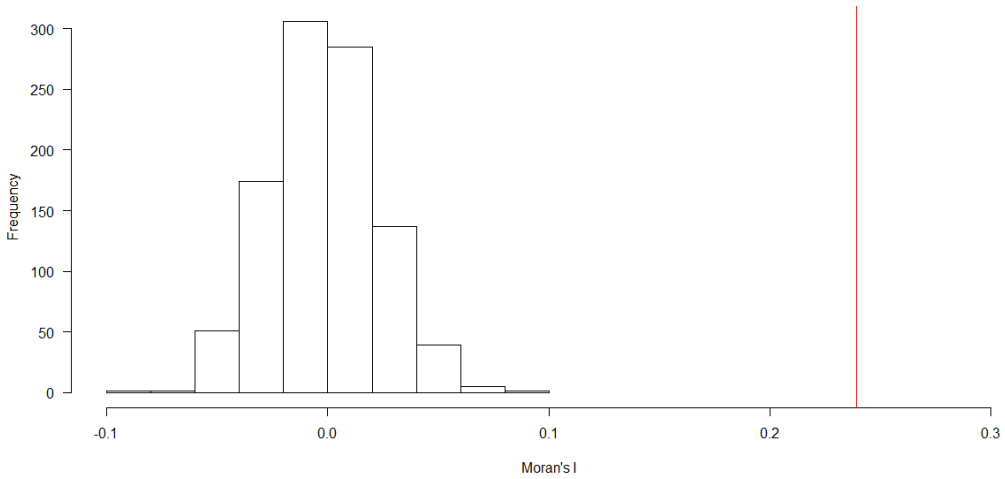


FIGURE 31: Distribution simulées des Indices de Moran et Indice observé

Nous avons en noir la distribution de l'indice sur les 1000 permutations et en rouge la valeur réelle de l'indice. On observe que les indices oscillent entre -0.05 et 0.05. Notre valeur étant à 0.24, il est fort probable qu'elle soit significative. La simulation

```
        Monte-Carlo simulation of Moran I

data:  ComteMoran$Nombre
weights: w
number of simulations + 1: 10001

statistic = 0.26259, observed rank = 10001, p-value = 9.999e-05
alternative hypothesis: greater
```

FIGURE 32: Simulation de Monte-Carlo de l'indice de Moran

de Monte-Carlo nous montre que les 10 000 Indices de Moran simulés sont inférieur à notre Indice observé. La p-value est égale à 1/10001 et on rejette l'hypothèse d'absence d'autocorrélation spatiale.

5.3 Régression spatiale

5.3.1 Motivations

Nous avons vu précédemment que la corrélation spatiale de la sinistralité était une des raisons qui a poussée les assureurs privés américains dans les années 20 à quitter le marché de l'assurance inondation. Nous proposons dans cette sous-partie des méthodes pour modéliser cette autocorrélation à partir de modèles de régression spatiale.

Ce que l'on considère comme la première loi de géographie, énoncée par Waldo Tobler, est la suivante : "Tout interagit avec tout, mais deux objets proches ont plus de chance de le faire que deux objets éloignés". Nous pouvons établir à priori l'hypothèse que si une zone est touchée par une inondation, en raison de la présence d'un facteur de risque (fleuve, rivière ou présence d'un littoral) alors les zones voisines ont de fortes chances d'avoir le même facteur de risque et donc d'avoir une sinistralité importante.

Lors d'une régression classique, nous supposons que les observations sont indépendantes entre elles. Les modèles issus de l'économétrie spatiale ont pour objectif de tenir compte de la corrélation entre les observations selon leur degré de proximité. Nous présentons ces modèles et les raisons qui nous poussent à déterminer lequel est le plus adapté selon le phénomène modélisé.

5.3.2 Matrice de voisinage

La première étape dans l'analyse de la corrélation spatiale entre différentes observations est la définition de le relation de voisinage entre les observations, à savoir définir quelles sont les observations qui sont voisines ou non. Nous nous appuierons pour cette partie

Nous pouvons définir de manière mathématique le phénomène modélisé (Insee, 2018) :

"Les relations spatiales $\mathcal{B}$ sont un sous ensemble du produit Cartésien $\mathcal{R}^2 \times \mathcal{R}^2 = (i, j) : i \in \mathcal{R}^2, j \in \mathcal{R}^2$ des couples (i, j) d'objets spatiaux, c'est-à-dire l'ensemble des couples (i, j) tels que i et j soient tous deux des objets spatiaux identifiés par leurs coordonnées géographiques et que (i, j) soit différent de (i, j). Un objet spatial ne peut être relié à lui-même (principe d'irreflexibilité : $(i, i) \notin \mathcal{B}$. De plus si $(i, j) \subseteq \mathcal{B}$ et si $(j, i) \subseteq \mathcal{B}$ pour tout couple d'objets spatiaux, les relations spatiales sont dites symétriques (Tiefelsdorf 1998)"

À la différence des relations temporelles qui ne s'articulent qu'autour d'un axe temporel, les relations spatiales sont multidirectionnelles et multilatérales (Insee, 2018). Afin de codifier la structure de voisinage, nous passons par une matrice de voisinage : Soit S une surface divisée en n zones mutuellement exclusives. Nous attribuons à chaque zone un point de référence : son centroïde. Les relations spatiales

peuvent être matérialisées par un graphe de voisinage qui relie les zones définies comme voisines, ou une matrice qui contient les coordonnées géographiques des points de référence. Enfin, nous codons le graphe dans une matrice de voisinage W tel que :

$$w_{ij} = \begin{cases} 1 \; si \; i \; et \; j \; sont \; voisins \\ 0 \; sinon \end{cases}$$

Prenons l'exemple de Tiefelsdorf (1998). Soit une aire divisée en 5 zones dont leur centroïde a été déterminé.

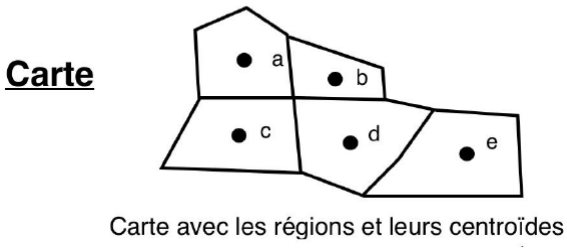


FIGURE 33: Exemple d'une division géographique d'un territoire, Insee 2018

Nous spécifions le graphe de voisinage, puis nous déterminons la matrice de voisinage W.

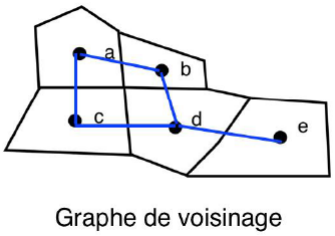


FIGURE 34: Exemple d'un graphe de voisinage d'un territoire, Insee 2018

Notre matrice de voisinage W s'écrit :

$$W = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Définition du voisinage

Le cœur du sujet est donc de définir la spécification de voisinage. Plusieurs méthodes sont possibles, premièrement la définition du voisinage par la distance entre les centroïdes :

— La triangulation de Delaunay qui relie les points sous forme de triangle afin que l'angle minimal des triangles soit maximisé.

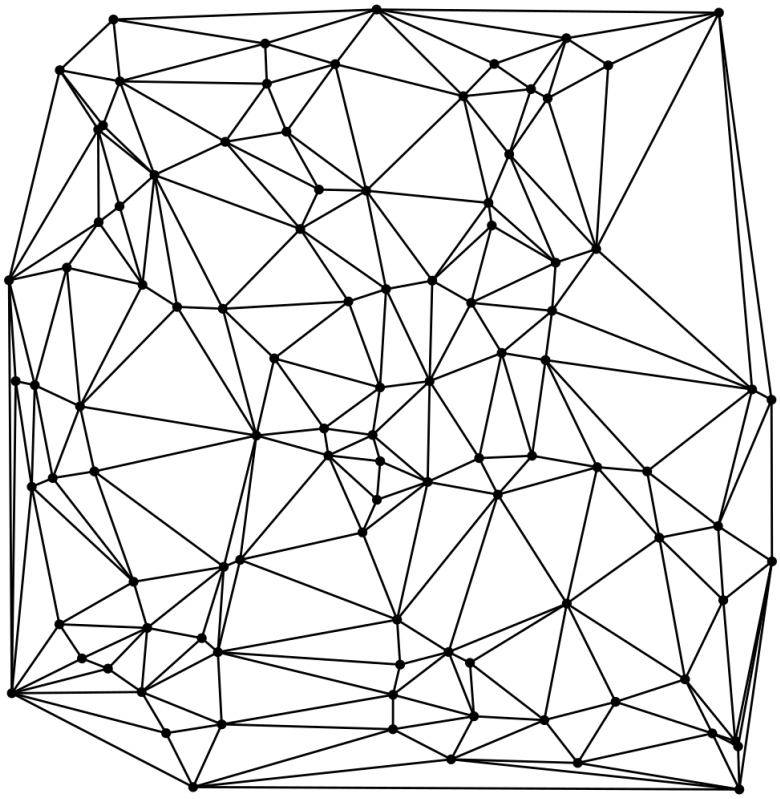


Figure 35: Triangulation de Delaunay. Source : Wikimedia Commons

— Le graphe de la sphère d'influence passe par la définition du "cercle du voisin le plus proche" d'un point a qui est le plus grand cercle centré en a qui ne contient pas d'autre point que a. Deux points sont considérés comme voisins si leur cercles se coupent.

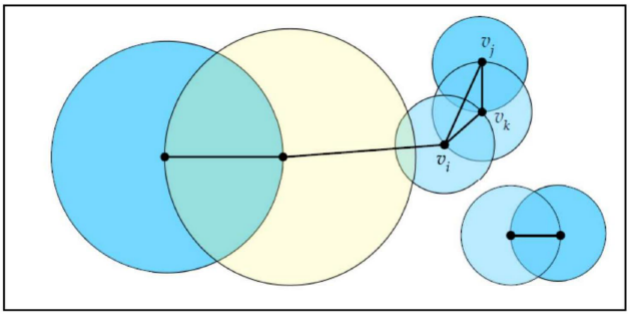


FIGURE 36: Graphe de la sphère d'influence. INSEE 2018

— Le graphe de Gabriel relie deux points a et b si et seulement si tous les autres points sont en dehors du cercle de diamètre [a,b].

Une autre méthode, qui s'appuie sur la contiguïté, est possible pour définir les zones voisines. Cette méthode est utilisée lorsque l'on est en présence de données qui représentent une partition d'un territoire, par exemple des données par département, ou dans notre cas par comté.

Cette méthode est utilisée en raison de l'ambiguïté que peut introduire le calcul de la distance entre les centroïdes. Prenons l'exemple de trois zones distinctes R_1, R_2 et R_3. On peut établir que les R_2 et R_3 ne sont pas voisines, pourtant les centroïdes des trois zones sont équidistants entre eux.

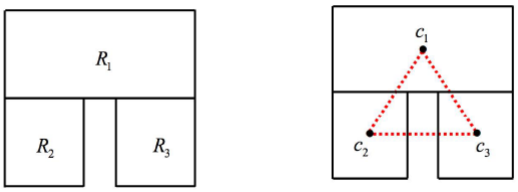


FIGURE 37: Exemple de centroïdes équidistants, Insee 2018

Deux méthodes sont possibles pour définir le voisinage par contiguïté :

— La contiguïté Rook : les zones doivent posséder au moins un segment de frontière en commun pour être considérées comme voisines.
— La contiguïté Queen : les zones doivent posséder au moins un point en commun pour être considérées comme voisines.

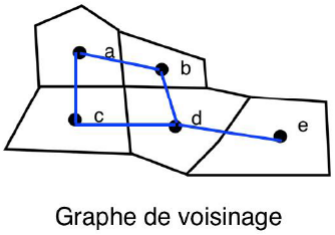


FIGURE 38: Exemple d'un graphe de voisinage d'un territoire

Notre matrice de voisinage au sens Rook W s'écrit :

$$W = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Notre matrice de voisinage au sens Queen W' s'écrit :

$$W = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Nous présentons ici la définition du voisinage des comtés de l'État de l'Iowa au sens Rook et Queen.

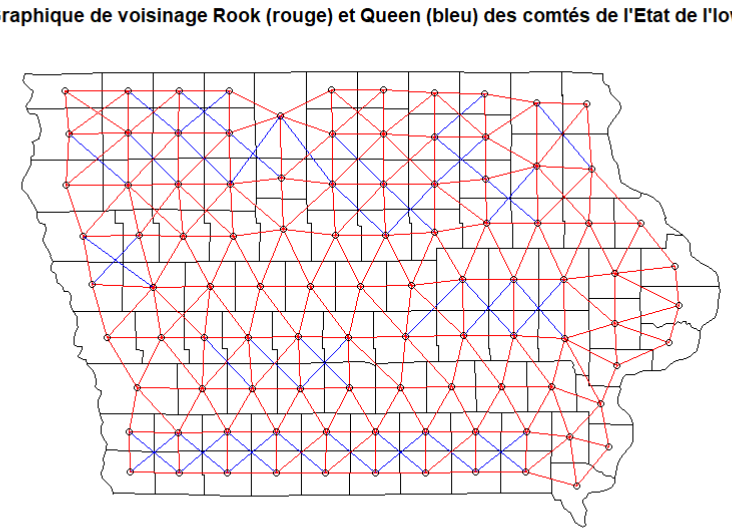


FIGURE 39: Relation de voisinage Rook et Queen des comtés de l'Etat de l'Iowa

Définition de la matrice de poids

Après avoir défini la matrice de voisinage selon une méthode choisie, il convient de transformer cette matrice en une matrice de poids. Cette matrice de poids est définie par Anselin et al. (1988) , comme "l'expression formelle de la dépendance spatiale entre observations." Le passage de la matrice de voisinage à la matrice de poids passe par une normalisation de cette première. Nous définissons le degré de liaison comme la somme des poids des voisins d'une zone. L'absence de normalisation de la matrice de poids entraîne de l'hétérogénéité entre les zones, car le degré de liaison dépendra du nombre de ses voisins.

Nous pouvons définir quatre types de normalisation :

- La normalisation en ligne : pour chaque zone, le poids accordé à chaque voisin est divisé par la somme des poids de ses voisins tel que $\sum_{i=1}^{j} w_{i,j} = 1$. Cette normalisation permet d'interpréter plus facilement la matrice de poids car $\sum_{j=1}^{j} w_{i,j}x_j$ représente tout simplement la moyenne de la variable d'observation x sur les voisins de la zone i.

- La normalisation globale : les poids sont standardisés afin que la somme de tous les poids soit égale au nombre total d'entités. C'est à dire que les poids sont multipliés par $\frac{n}{\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}}$.

- La normalisation uniforme : les poids sont standardisés afin que la somme de tous les poids soit égale à 1 : $\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij} = 1$

- La normalisation par stabilisation de variance : on définit un vecteur q tel que $q = \left(\sqrt{\sum_{j=1}^{n} w^2{}_{1j}}, \sqrt{\sum_{j=1}^{n} w^2{}_{2j}}..., \sqrt{\sum_{j=1}^{n} w^2{}_{nj}}\right)^T$. On définit la matrice $S* = diag(q)^{-1}W^4$ et $Q = \sum_{j=1}^{n}\sum_{i=1}^{n} s*_{ij}$. On définit la matrice de poids normalisée $S = \frac{n}{Q}S*$

Chaque méthode de normalisation entraîne des conséquences différentes. La normalisation en ligne donne un poids plus important aux observations situées en bordure de la surface étudiée, qui ont un faible nombre de voisins. La normalisation par stabilisation de la variance permet de réduire l'hétérogénéité dans les poids entraînée par les différences de nombre de voisins entre chaque zone. La normalisation uniforme permet aux observations présentes au centre de la surface étudiée, qui ont un nombre important de voisins, d'être plus influencées par ses voisins que ne le sont les observations situées en bordure.

La somme totale des éléments de la matrice W est toujours égale à n, peu importe le choix de la méthode de normalisation. Cela a pour conséquence que les statistiques d'autocorrélation spatiale sont comparables entre elles.

Le choix de la matrice de poids, et son impact dans les modèles d'économétrie spatiale font débat : "Le choix des poids est souvent arbitraire [...] et le résultat des études varie considérablement en fonction de la définition des poids spatiaux" (Bhattacharjee et al., 2005). Cela impliquerait qu'un mauvais choix de matrice de poids peut amener à de fausses conclusions. LeSage et al. (2010), au contraire, estiment que les différences entre les matrices de poids n'ont pas une influence significative sur les résultats.

Dans notre étude, nous choisiront la normalisation uniforme en ligne, pour des raisons d'interprétation des résultats, comme nous l'avons mentionné précédemment.

5.3.3 Modèles de régression spatiale

Motivations

Lorsque l'on étudie des données spatiales, on peut très souvent observer de l'autocorrélation spatiale des résidus, c'est-à-dire une dépendance entre des observations voisines. Ce phénomène peut remettre en cause l'utilisation de modèle linéaire classique : Dans le cadre de ce modèle, nous cherchons à estimer le phénomène observé par l'équation $Y = X\beta + \epsilon$ avec $Y \ \mathcal{N}_n(X\beta, \Sigma)$. S'il n'y a pas de dépendance entre les observations, alors $\Sigma = \sigma^2 I_n$ et nous estimons nos paramètres par la méthode des moindres carrés. S'il y en a alors Σ n'est plus une matrice diagonale. Dans le cas où il y a de la dépendance, spatiale dans notre cas, entre les observations alors la régression linéaire classique détermine des estimations biaisées et inconsistantes des paramètres. Les tests de Fisher et de Student que nous utilisons pour valider le modèle, ne sont plus valables dans ce cas.

Liste des modèles

Nous reprenons la présentation de Floch et Le Saoult (Insee, 2018) de la classification des modèles faites par Elhorst, en soulignant les trois types d'interactions spatiales issus des premiers modèles de Manksi.

- L'interaction endogène, lorsque l'observation d'une zone géographique dépend de l'observation des zones voisines.
- L'interaction exogène, lorsque l'observation d'une zone géographique dépend des caractéristiques des zones voisines.
- Une corrélation spatiale liée à des caractéristiques similaires non observées.

Le modèle de Manski s'écrit :

$$Y = \rho WY + \alpha 1_N + X\beta + WX\theta + u$$

avec $u = \lambda Wu + \epsilon$

— WY représente l'interaction endogène entre la variable d'intérêt quantifiée par le coefficient ρ, appelé coefficient d'autorégressif spatial.

— WX représente l'interaction exogène, quantifié par le vecteur θ (dont la taille est égale au nombre de variables explicatives)

— Wu représente l'interaction entre les erreurs du modèle, quantifiée par le coefficient λ, appelé coefficient d'autocorrélation spatiale.

Les auteurs rappellent que le modèle de Manski n'est pas identifiable sous cette écriture : l'estimation des paramètres θ, λ, ρ et β en même temps est impossible. L'exemple d'illustration par les auteurs est celui de l'analyse au sein d'une classe de l'autocorrélation spatiale des résultats des élèves. Il est supposé que les mauvais résultats d'une classe s'expliquent par la composition sociale de la classe, ce qui représente l'interaction exogène et la qualité des professeurs (caractéristique non observée). On observera une forte corrélation au sein de la classe des résultats des élèves sans que cela signifie que les résultats des élèves ont un effet sur leurs voisins, ce qui correspond à l'interaction endogène.

Pour que le modèle soit identifiable, plusieurs solutions sont possibles. La première est de définir des matrices de voisinage différentes pour chaque interaction spatiale. L'autre solution est d'enlever une des interactions du modèle, c'est-à-dire supposer qu'un des paramètres (ρ,θ,λ) est nul.

Le modèle de Manski impose certaines conditions : les matrices $I - \rho W$ et $I - \lambda W$ doivent être inversibles. L'usage de matrice de contiguïté ou de distance inverse permet de respecter ces conditions. On suppose également que $|\rho| < 1$ et $|\lambda| < 1$.

Les trois types de modèles qui découlent du modèle de Manski de l'hypothèse de nullité d'un des coefficients sont les suivants :

— Le modèle SDEM (Spatial Durbin Error Model) dans le cas où $\rho = 0$. On suppose que l'interaction endogène est nulle et que l'interaction exogène est plus importante dans le modèle à estimer.

— Le modèle de Kelejian-Prucha dans le cas où $\theta = 0$. Ce modèle comporte le défaut d'avoir des estimateurs biaisés et non convergents s'il y a présence d'interaction exogènes (Lesage, 2010) en raison de biais par variables omises.

— Le modèle spatial de Durbin (Spatial Durbin Model) dans le cas où $\lambda = 0$. Dans ce cas, les estimateurs ne sont pas biaisés et le modèle est plus robuste à une mauvaise spécification (Insee, 2018).

Les modèles de Kelejian-Prucha et Spatial Durbin Model comportent les cas particuliers du modèle SAR (Spatial Autoregression) et SEM (Spatial Error Model où $Y = X\beta + u$ avec $u = \lambda W u + \epsilon$)

— Le modèle SAR s'écrit en posant $\lambda = \theta = 0$. Dans ce cas le modèle s'écrit $Y = \rho WY + X\beta + \epsilon$

— Le modèle SEM s'écrit en posant $\theta = -\rho\beta$ (hypothèse appelée de facteur commun). Dans ce cas le modèle s'écrit $Y = X\beta + \rho W(Y - X\beta) + \epsilon$. En posant $u = Y - X\beta$, on obtient le modèle $Y = X\beta + u$ avec $u = \lambda$

On note également le modèle SLX (Spatial Lag X), en posant $\lambda = \rho = 0$

Choix des modèles

Une fois supposée la normalité des résidus $\epsilon \ \mathcal{N}(0, \sigma^2)$ et la matrice de voisinage est connue, les modèles sont estimés par maximum de vraisemblance. Deux méthodes sont possibles généralement pour choisir le modèle le plus adapté.

La première méthode est dite *bottom-up* (approche ascendante) : on utilise d'abord un modèle linéaire classique. On utilise ensuite des tests du multiplicateur de Lagrange pour déterminer quel modèle choisir entre le SAR, SEM ou un modèle non spatial.

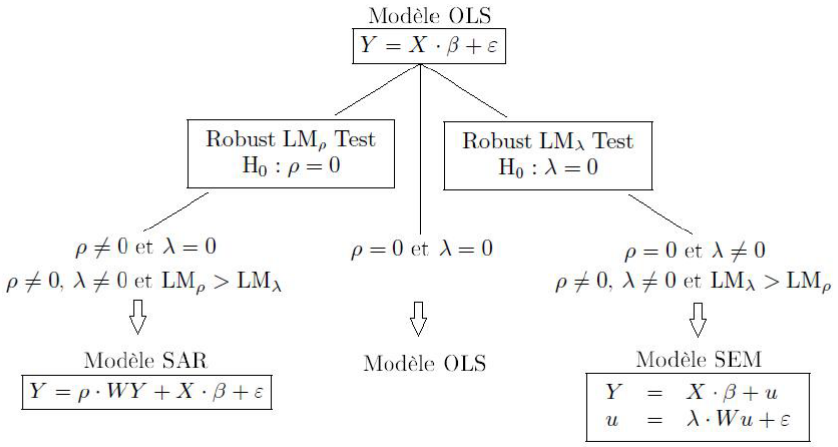


FIGURE 40: Méthode bottom-up de sélection de modèles spatiaux, Insee 2018

La deuxième méthode est dite *top-down* (approche descendante) : on utilise d'abord le modèle spatial de Durbin. On détermine ensuite quel modèle tester grâce à un test du rapport de vraisemblance.

Une autre approche intermédiaire a été proposée par Elhorst. Elle est similaire à l'approche ascendante pour la première étape, mais en cas d'interactions spatiales, on étudie le modèle spatial de Durbin au lieu de choisir entre le modèle SAR ou SEM. On utilise ensuite des tests de rapport de vraisemblance pour choisir le modèle le plus adapté.

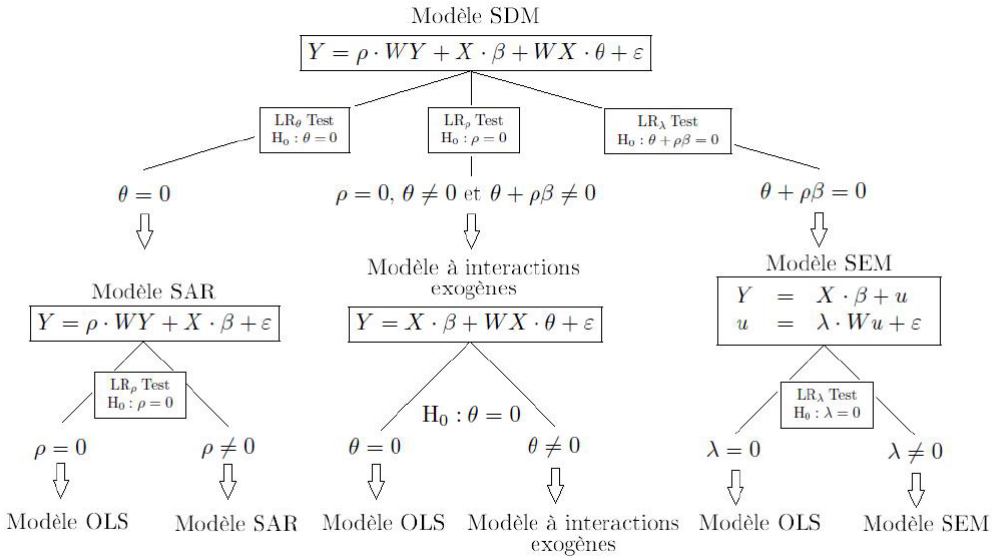


FIGURE 41: Méthode top-down de sélection de modèles spatiaux, Insee 2018

Interprétation des résultats

L'interprétation des résultats dans le cadre d'une régression spatiale peut être différente que celle dans le cadre d'une régression classique. En effet, la prise en compte de la dépendance entre les zones voisines fait que la variation d'une variable explicative sur une zone va impacter à la fois la variable réponse de cette zone mais aussi la variable réponse des zones voisines. Il y a alors un effet multiplicateur global qui va impacter l'ensemble de l'échantillon. Cependant dans le cadre d'une autocorrélation spatiale des erreurs, l'interprétation des résultats est identique à celle d'une régression classique. La variation d'une variable explicative sur une zone va impacter sa variable réponse et indirectement les zones voisines, sans que cet effet soit démultiplié (Insee, 2018).

Nous pouvons expliciter mathématiquement ces interactions en reprenant l'exemple des auteurs (Insee, 2018) dans le cadre du modèle SAR ($Y = \rho WY + X\beta + \epsilon$). Dans ce cas

$$Y = (1 - \rho W)^{-1} X\beta + (1 - \rho W)^{-1}\epsilon$$

$= \sum_{r=1}^{k} (1 - \rho W)^{-1}\beta_r X_r + (1 - \rho W)^{-1}\epsilon$

$= \sum_{r=1}^{k} S_r(W) X_r^{-1}\beta_r X_r + (1 - \rho W)^{-1}\epsilon$

avec $S_r(W)_{ii} = (1 - \rho W)^{-1}$

La variable prédite se calcule donc $\hat{y} = (1 - \rho W)^{-1} X\hat{\beta}$ alors que dans un modèle linéaire classique la variable prédite se calcule $\hat{y} = X\hat{\beta}$

Dans le cas de la régression spatiale, l'effet marginal d'une variation de la variable X_r pour une zone i n'est pas β_r mais $(S_r(W)_{ii})$.

L'effet marginal est différent pour chaque zone. Les éléments de la diagonale de la matrice S_r représentent les effets directs d'une variation de la variable X_r sur la zone. Les autres éléments de la matrice S_r représentent les effets indirects, à savoir l'impact de la modification de la variable X_r d'une zone sur les zones voisines. On détermine pour l'ensemble du territoire d'étude, les effets directs et indirects en calculant la moyenne de ces effets (Lesage 2010).

- L'effet direct moyen est la moyenne des éléments diagonaux de la matrice S_r : Effet direct moyen$= 1/n * trace(S_r)$
- L'effet total moyen est la moyenne de l'ensemble des éléments de la matrice $S_r : 1/n * \sum_i[\sum_j S_r(W)_{ik}]$. Deux interprétation de cette moyenne sont possibles. Cet effet peut représenter la moyenne de tous les effets sur une zone i d'une variation de la variable X_r, ou bien comme la moyenne de tous les effets d'une variation de la variable X_r dans une zone i sur l'ensemble des zones.
- L'effet indirect moyen se calcule grâce à la différence entre l'effet total moyen et l'effet direct moyen.

5.3.4 Application des modèles

Nous choisissons d'étudier ces modèles spatiaux en les appliquant au taux de destruction à l'échelle des comtés sur notre sélection d'États. Nous cherchons à voir s'il y a une autocorrélation spatiale entre les comtés en ce qui concerne leur taux de destruction, comment la mettre en évidence et voir comment les modèles spatiaux peuvent améliorer les modèles classiques.

Nous agrégeons les données au niveau des comtés en prenant la somme des sinistres divisée par la somme de la valeur assurée pour obtenir notre taux de destruction. Les variables explicatives sont agrégées en prenant la proportion des modalités de références : nous calculons le pourcentage de maisons se trouvant en zone à risque ou à faible risque, de maisons ayant un étage et ainsi de suite pour toutes nos variables explicatives.

Choix de la matrice de voisinage

Étant donné que nous sommes en présence de données surfaciques, nous choisissons pouvons choisir de définir nos voisins selon la méthode Queen ou Rook. Afin de limiter le nombre de voisins sélectionné par zone, nous choisissons la méthode Rook. Nous montrons en annexe les résultats des modèles avec la méthode Queen, qui ne diffèrent pas sensiblement de la méthode Rook.

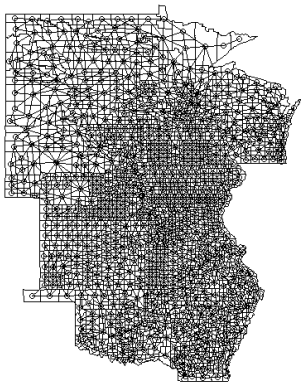


FIGURE 42: Relations de voisinage de type Queen entre les comtés

Application des modèles classiques

Nous modélisons dans un premier temps le taux de destruction par un modèle linéaire classique, en composant le taux de destruction par trois fonctions : la fonction identité, la fonction log et la fonction racine. Nous étudions les résidus de la régression :

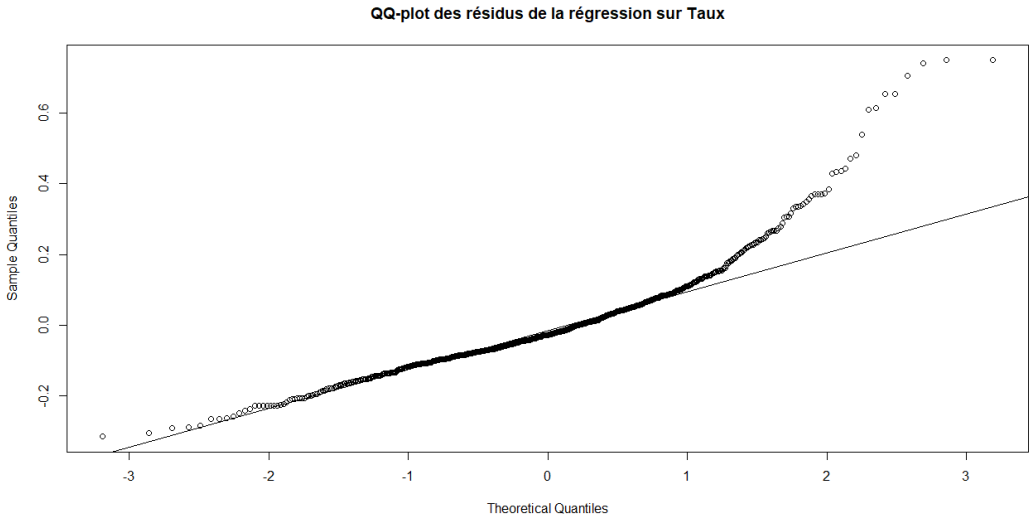


FIGURE 43: Résidus de la régression linéaire classique sur le taux de destruction

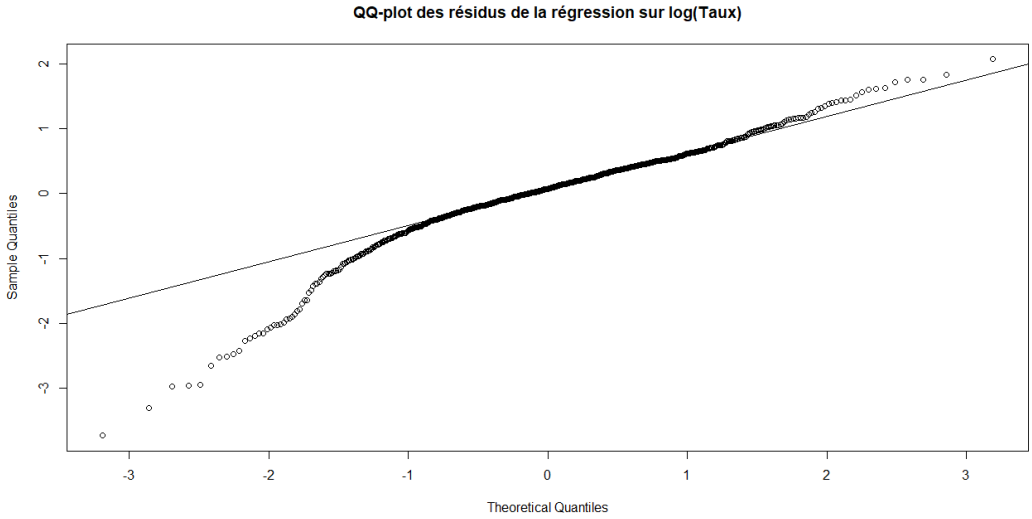


FIGURE 44: Résidus de la régression linéaire classique sur le logarithme du taux de destruction

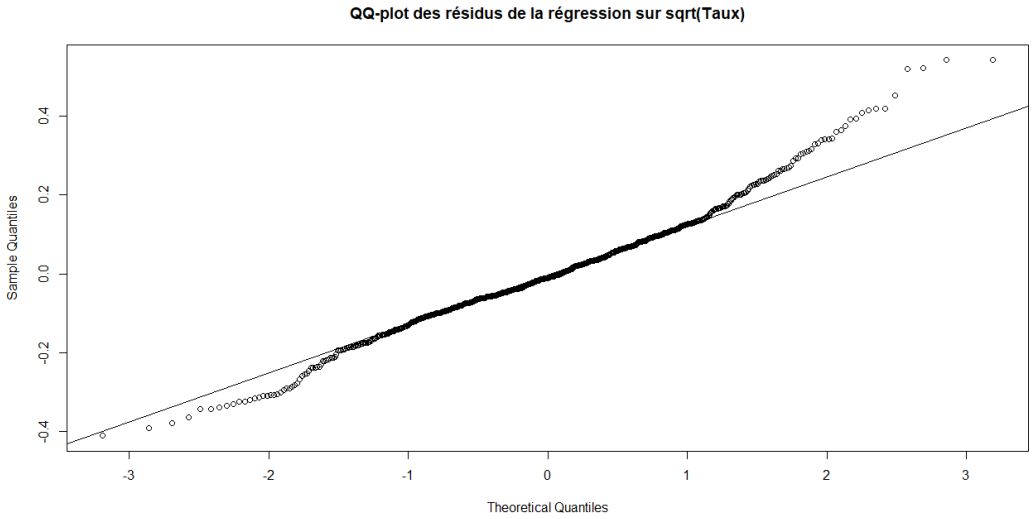


FIGURE 45: Résidus de la régression linéaire classique sur la racine du taux de destruction

Nous voyons que l'ajustement des résidus à une loi normale est assez mauvais sur les valeurs extrêmes. En observant le diagramme en boite des taux observés nous voyons détectons la présence de données exceptionnelles.

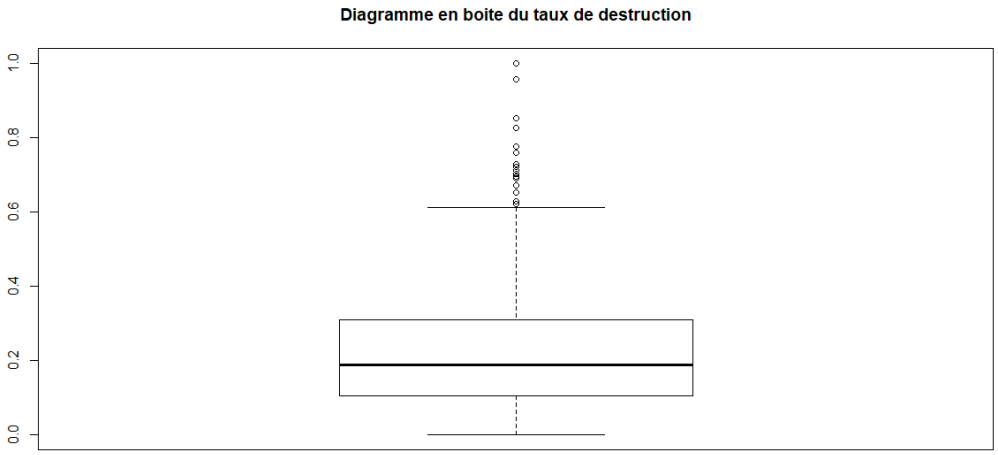


FIGURE 46: Diagramme en boite du taux de destruction par comté

Une méthode pour traiter ces valeurs extrêmes est la Winsorisation. Elle consiste à attribuer la valeur d'un certain quantile à toutes les valeurs dépassant ce quantile. Dans notre cas, nous prenons le quantile à 95 %. Nous obtenons le taux de destruction 0.53 %. Les 5 % des observations dépassant le taux de 53 % se voient attribuer ce taux. La modification de données doit toujours être faite avec précaution. Comme nous sommes dans l'étude de modèles spatiaux, nous souhaitons conserver l'ensemble de nos observations étant donné les relations de voisinage. De plus, le modèle linéaire classique est le point de départ de notre étude sur les modèles spatiaux. Nous voyons que les valeurs extrêmes sont mal ajustées par le modèle gaussien. Nous décidons de continuer avec ce modèle et de donner moins d'importance aux valeurs extrêmes.

Nous relançons notre régression et nous obtenons les résultats suivants :

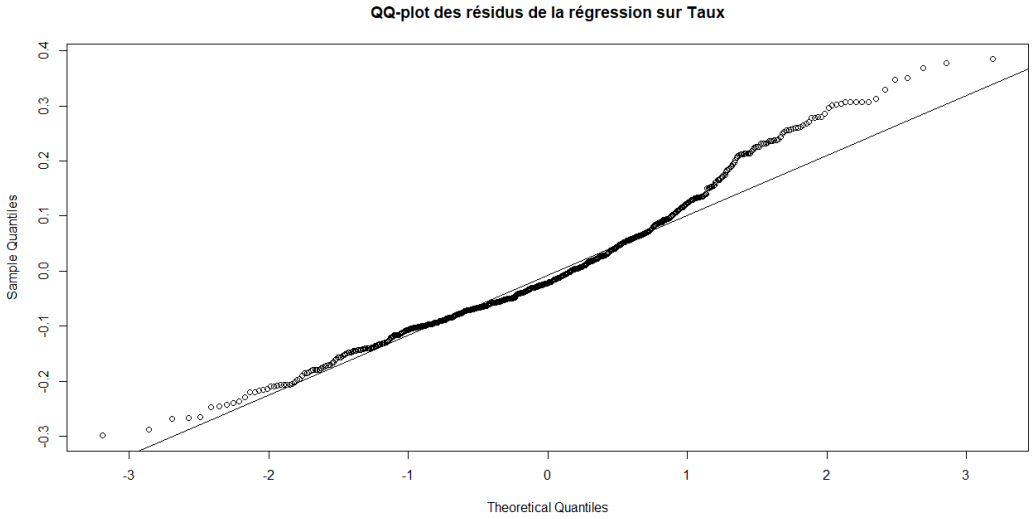


FIGURE 47: Résidus de la régression linéaire classique sur le taux de destruction

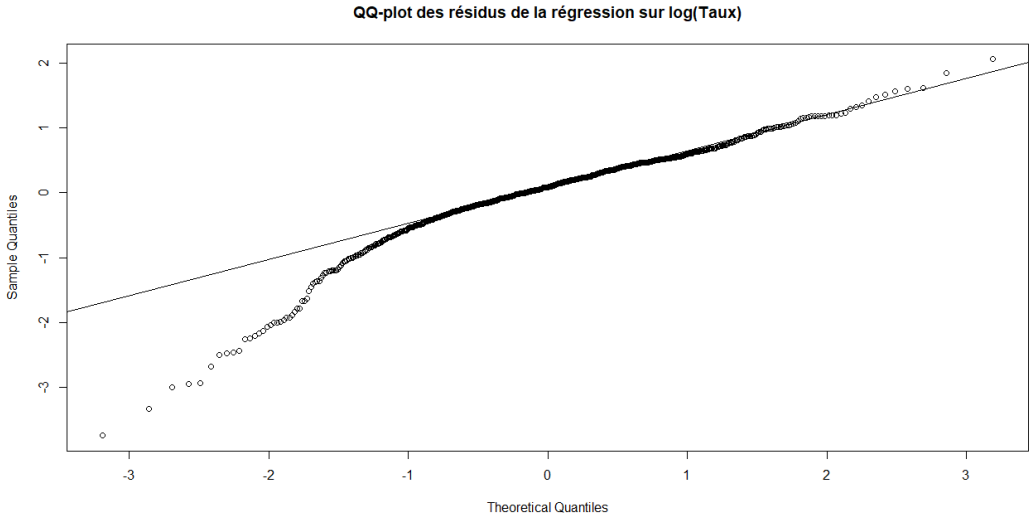


FIGURE 48: Résidus de la régression linéaire classique sur le logarithme du taux de destruction

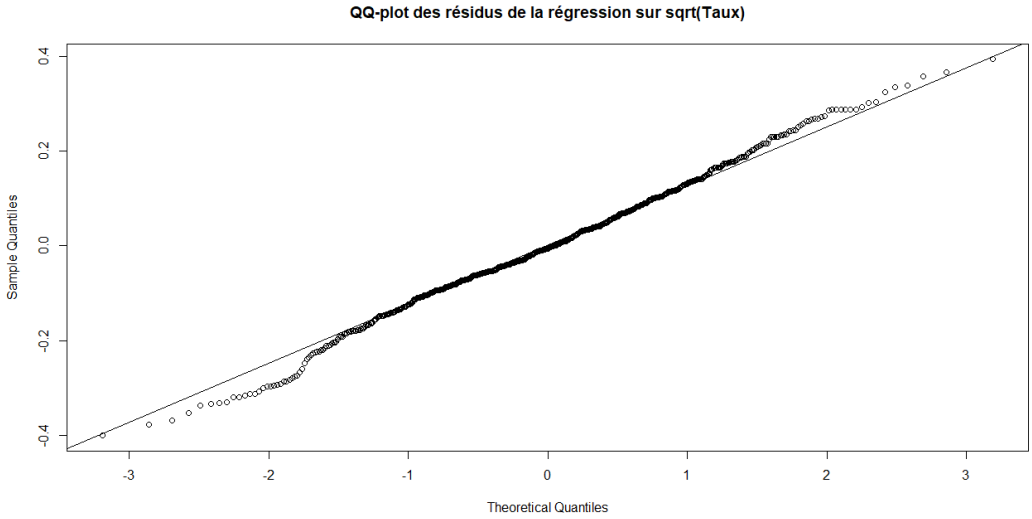


FIGURE 49: Résidus de la régression linéaire classique sur la racine du taux de destruction

Nous constatons que les résidus de la régression avec la fonction racine sont ceux qui sont le mieux ajustés à une loi normale. Nous continuons avec ce modèle.

Nous effectuons un test de Moran sur les résidus de la régression pour identifier une éventuelle corrélation spatiale qui nous amènerait à penser qu'un modèle spatial pourrait améliorer la vraisemblance du modèle.

Nous obtenons les résultats suivants :

Indice de Moran	0.104
P-Value	7.442e-06

Nous ne rejetons pas l'hypothèse d'autocorrélation. L'indice de Moran à 0.104 nous indique la présence d'une légère autocorrélation positive.

Après avoir identifié l'autocorrélation spatiale des résidus de la régression linéaire, nous cherchons à savoir quel modèle spatial utiliser pour ajuster notre régression. Pour cela nous utilisons le test du multiplicateur de Lagrange, qui nous permet d'arbitrer entre un modèle SAR (Modèle spatial auto-régressif) ou SEM (Modèle d'erreurs spatiales).

Le test du multiplicateur de Lagrange utilise les statistiques suivantes :

- Pour un modèle SEM :
$$LM_{SEM} = \frac{(\hat{\epsilon}'W\hat{\epsilon})^2}{T\hat{\sigma}^4}$$
avec $T = tr((W' + W)W)$ et $\hat{\epsilon}$ et $\hat{\sigma}$ sont les estimateurs de ϵ et σ sous H0. Sous l'hypothèse H0 on a :
$$LM \longrightarrow \chi^2(1)$$
- Pour le modèle SAR :
$$LM_{SAR} = \frac{(\hat{\epsilon}WY)^2}{\hat{T}\hat{\sigma}^4}$$
avec $\hat{\epsilon}$ et $\hat{\sigma}$ sont les estimateurs de ϵ et σ sous H0 et $T = ((WX\hat{\beta}(I - X(X'X)^{-1}X')(WX\beta) + T\hat{\sigma^2}/\hat{\sigma}^2$ Sous l'hypothèse H0 on a :
$$LM_{SAR} \longrightarrow \chi^2(1)$$

Les deux tests que nous venons de présenter permettent de savoir si nous devons prendre en comte dans le modèle l'autocorrélation des erreurs ou, l'autocorrélation de la variable réponse. Ces deux tests ne tiennent pas compte du fait qu'il pourrait y avoir d'autres formes d'autocorrélations. Pour cela nous pouvons tester les formes robustes de ces tests :

— La forme robuste du test du modèle SEM nous permet de savoir si nous devons prendre en compte dans le modèle l'autocorrélation des erreurs, sachant qu'il pourrait y avoir une autocorrélation de la variable à expliquer qui n'est pas prise en compte dans le modèle.
— La forme robuste du test du modèle SAR nous permet de savoir si nous devons prendre en compte dans le modèle l'autocorrélation des résidus, sachant qu'il pourrait y avoir une autocorrélation des erreurs qui n'est pas prise en compte dans le modèle.

Nous obtenons les résultats suivants :

Modèle	**p-value**
LM Error	1.942e-05
RLM Error	0.4693
LM Lag	2.85e-06
RLM Lag	0.04057

Les résultats des tests de Lagrange nous amène à privilégier le modèle SAR par rapport au modèle SEM.

Le modèle SAR estimé nous donne les résultats suivants :

```
Residuals:
       Min         1Q     Median         3Q        Max
-0.4059472 -0.0757893 -0.0065793  0.0849522  0.3866617

Type: lag
Regions with no neighbours included:
 284 318 543 570 576 745
Coefficients: (asymptotic standard errors)
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept)  0.357324   0.024122  14.8134 < 2.2e-16
‘Sous Sol 1‘ -0.305697   0.024106 -12.6812 < 2.2e-16
‘Sous Sol 2‘ -0.070212   0.024612  -2.8528  0.004334
A             0.106121   0.019747   5.3739 7.703e-08
AE            0.119462   0.019321   6.1829 6.292e-10

Rho: 0.19865, LR test value: 19.746, p-value: 8.8445e-06
Asymptotic standard error: 0.0447
    z-value: 4.444, p-value: 8.8305e-06
Wald statistic: 19.749, p-value: 8.8305e-06

Log likelihood: 424.6055 for lag model
ML residual variance (sigma squared): 0.017259,
(sigma: 0.13138)
Number of observations: 700
Number of parameters estimated: 7
```

```
AIC: -835.21, (AIC for lm: -817.47)
LM test for residual autocorrelation
test value: 0.092413, p-value: 0.76113
```

Le coefficient Rho, qui représente la mesure de dépendance spatiale dans notre modèle, est positif et est significatif grâce aux résultats de deux tests : un calculé à partir de la matrice de variance asymptotique et l'autre grâce au rapport du test de vraisemblance. Cela signifie que les valeurs des taux de destructions voisins a un effet positif sur le taux de destruction de chaque zone : pour chaque comté, si le taux de ces voisins augmente, alors le taux de destruction du comté augmentera.

Nous voyons que nous obtenons un AIC plus faible dans le modèle SAR que dans le modèle linéaire classique. Une mesure pour déterminer si le gain en réduction en AIC est intéressant est la suivante :

$$\Delta = \frac{(logLik(M2) - logLik(M1))}{n} > 0.001$$

Dans notre cas, nous avons Δ=0.028. La réduction de l'AIC s'avère intéressante. De plus, nous constatons que l'hypothèse d'autocorrélation des résidus de la régression SAR peut être rejetée. Nous n'avons plus d'autocorrélation résiduelle puisque l'ensemble de l'autocorrélation a été captée par la variable supplémantaire introduite par le modèle SAR Nous pouvons tracer un diagramme de Moran pour remarquer les différences entre les résidus du modèle linéaire classique et celui du modèle SAR.

Comme nous l'avons vu précédemment, l'interprétation des résultats des coefficients de la régression. Dans le modèle spatial, nous devons nous intéresser aux effets directs et indirects des variables explicatives. Les effets directs, qui mesurent l'impact de la variation de la variable explicative sur la variable réponse, et les effets indrects qui mesurent l'impact sur la variable réponse de la variation de la variable explicative de ses voisins.

```
Impact measures (lag, exact):
                  Direct    Indirect       Total
'Sous Sol 1' -0.30832688 -0.07249968 -0.38082656
'Sous Sol 2' -0.07081599 -0.01665160 -0.08746759
A             0.10703344  0.02516774  0.13220117
AE            0.12048957  0.02833180  0.14882137
```

Dans notre cas nous voyons que l'effet indirect est relativement peu élevé et a le même signe que l'effet direct pour chacune des variables explicatives. On peut donc comprendre que les types de sous-sol 1 et 2 ont un effet négatif sur le taux destruction, tandis que plus la proportion de maison en zone inondable est élevée, plus le taux de destruction sera élevée.

Nous pouvons tracer le diagramme de Moran des résidus de la régression linéaire classique (Figure 50) et du modèle SAR (Figure 51) pour observer la captation par ce dernier de l'autocorrélation spatiale :

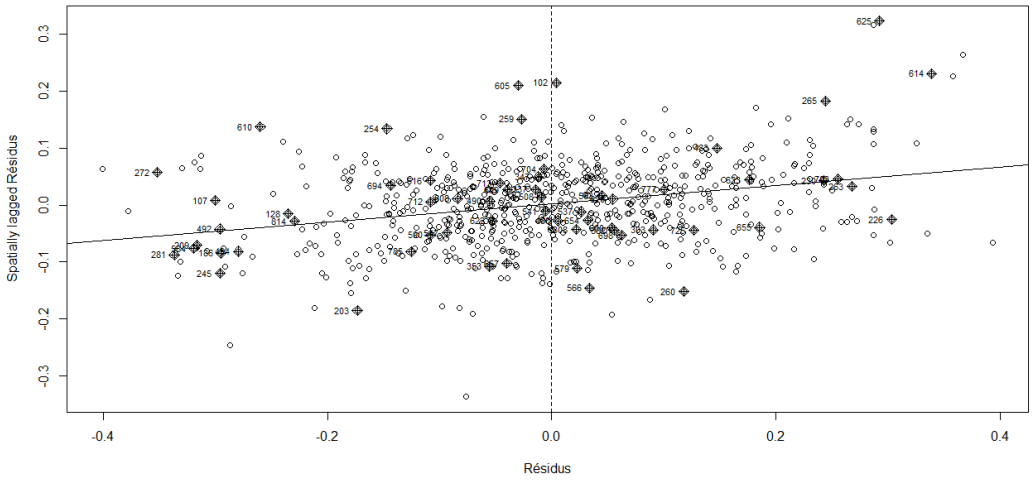


FIGURE 50: Diagramme de Moran des résidus de la régression linéaire classique

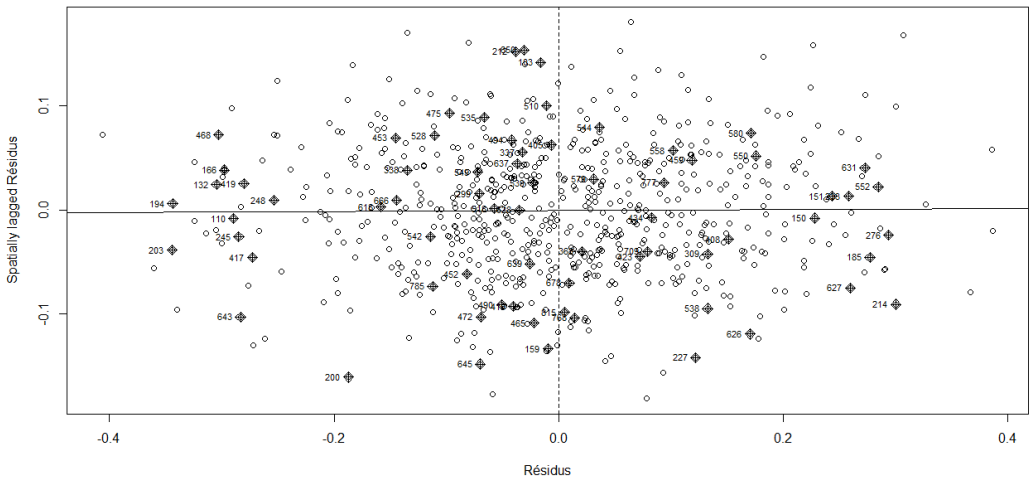


FIGURE 51: Diagramme de Moran des résidus du modèle SAR

Nous constatons que les résidus de la régression SAR ont une répartition plus aléatoire dans l'espace que les résidus issus de la régression linéaire classique.

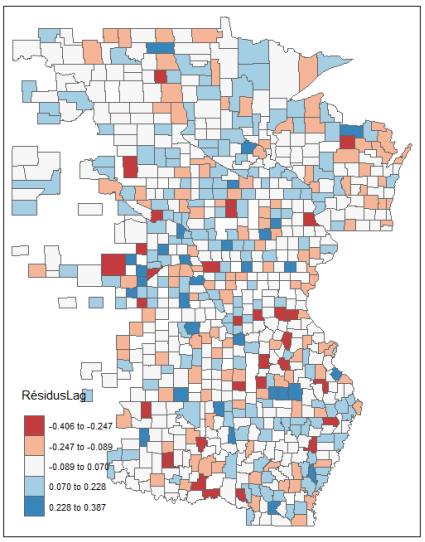


FIGURE 52: Résidus du modèle SAR

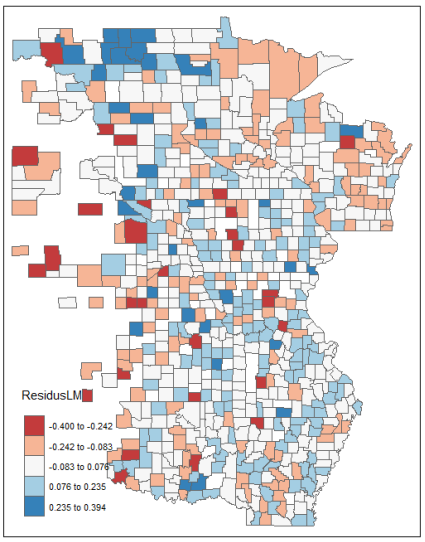


FIGURE 53: Résidus du modèle linéaire classique

5.4 Conclusion

Nous avons vu comment l'implémentation d'un modèle spatial permet d'améliorer la vraisemblance et de tenir compte de l'autocorrélation des observations, par rapport à un modèle linéaire classique. Néanmoins ces modèles peuvent être critiqués en raison de leur manque de robustesse selon la relation de voisinage choisie. De

plus, il peut y avoir un risque d'erreur "écologique", à savoir un découpage territoriale non pertinent pour l'étude du phénomène modélisé. Nous avons pris dans notre cas le découpage administratif des comtés, ce qui peut être remis en cause en raison de leur périmètre, de la variabilité du risque au sein d'un même comté et d'un découpage qui ne reflète pas le risque naturel. Il y a également un autre phénomène sous-jacent possible qui est celui de l'hétérogénéité spatiale, qui peut se manifester de deux manières :

- la variabilité spatiale des paramètres estimés, à savoir "l'absence de stabilité dans l'espace des comportements ou des relations économiques".
- l'hétéroscédasticité, à savoir la variabilité des variances des termes d'erreurs selon la localisation.

Il peut être difficile de distinguer ces phénomènes de l'autocorrélation spatiale rappelle Floch et Le Saoult (2018).

Conclusion

Nous avons vu que le risque inondation possède des caractéristiques particulières qui le rendent singuliers par rapport à d'autres risques. Les méthodes mises en œuvre pour améliorer la résilience du risque diffèrent selon les pays et pose la question du niveau de mutualisation acceptable pour les assurés, étant donné que le risque est très localisé. Dans cette perspective, nous avons étudié les manières de détecter un phénomène d'autocorrélation et l'implémentation de modèles spatiaux pour prendre en compte cette autocorrélation spatiale dans des modèles de régressions linéaires. Nous avons détaillé les étapes nécessaires pour implémenter ces modèles issus de l'économétrie spatiale ainsi que les faiblesses qu'ils peuvent comporter (éventuel manque de robustesse de la sélection de voisinage ou la définition arbitraire des zones). Nous avons pu améliorer la vraisemblance de notre modèle et tenir compte de la dépendance spatiale des observations en supprimant l'autocorrélation spatiale des résidus.

Ces modèles spatiaux ont de nombreux débouchés dans les sciences actuarielles, notamment dans le domaine de l'assurance paramétrique. Les données récoltés par des stations de mesures (pluie, vent, niveau des rivières) peuvent inclure de la dépendance spatiale. Muhammad et Lu (2020) ont étudié la modélisation d'un indice QMED, à savoir la médiane des maximum annuels du débit d'eau ayant une période de retour de 2 ans. Leurs travaux ont porté sur l'étude de 586 stations hydrométriques au Royaume-Uni en utilisant des données géologiques et météorologiques. Les auteurs ont montré que l'intégration d'un modèle spatial SEM (Spatial Error Model) permettait d'améliorer significativement les modèles existants.

D'autres travaux ont porté sur la sinistralité en assurance automobile (Tufvesson, J. Lindström et E. Lindström 2019). Les auteurs ont montré que l'intégration de la dépendance spatiale dans l'estimation de la fréquence des sinistres permettait d'améliorer leurs résultats par rapport à des modèles linéaires généralisés classiques.

Bibliographie

Anselin, Luc et Daniel A Griffith GRIFFITH (1988). *Do spatial effects really matter in regression analysis ?* Papers in Regional Science 65.1, p. 11–34.

Bhattacharjee, Arnab et Chris Jensen-Butler (2005). *Estimation of spatial weights matrix in a spatial error model, with an application to diffusion in housing demand.* CRIEFF Discussion Papers.

Carolyn Kousky and Leonard Shabman (2014) *How and Why the NFIP Differs from a Private Insurance Company*

Congressional Budget Office (2009). *The National Flood Insurance Program : Factors Affecting Actuarial Soundness*

Congressional Research Service. (2019). *Introduction to the National Flood Insurance Program*

Federal Emergency Mangement Administration (2013) *Technical Documentation of NFIP Actuarial Assumptions and Methods.*

Federal Emergency Management Administration. *NFIP Redacted Claims. NFIP Redacted Policies.*

https ://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims

Federal Emergency Management Agency, *NFIP CRS Fact Sheet.*

Francisco Cribari-Neto & Achim Zeilis. *Beta Regression in R.*

https ://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf

INSEE. *Manuel d'analyse spatiale.* 2018

LESAGE, James P et R Kelley PACE (2010). *The biggest myth in spatial econometrics.*

Laurence Reboul, cours de statistique spatiale, Aix-Marseille Université : http ://iml.univ-mrs.fr/ reboul/enseignement.html

Keating, A. et al. (2014), *Operationalizing Resilience against Natural Disaster Risk : Opportunities, Barriers, and a Way Forward*, Zurich Flood Resilience Alliance.

Kunreuther, *Reauthorizing the National Flood Insurance Program*, *Issues in science and technology*

Muhammad, M., Lu, Z. *Estimating the UK Index Flood : an Improved Spatial Flooding Analysis. Environ Model Assess 25*, 731–748 (2020).

National Research Council. (2015). *Affordability of National Flood Insurance Program Premiums : Report 1.* Washington, DC : The National Academies Press. https ://doi.org/10.17226/21709.

OECD (2016), *Financial Management of Flood Risk*, OECD Publishing, Paris. *http ://dx.doi.org/10.1787/9789264257689-en*

Oskar Tufvesson, Johan Lindström & Erik Lindström (2019) *Spatial statistical modelling of insurance risk : a spatial epidemiological approach to car insurance*, Scandinavian Actuarial Journal, 2019 :6, 508-522

UNISDR (2013), *2013 floods a "turning point"*, UNISDR, Geneva, 25 June, www.unisdr.org/archive/33693, accessed 6 January 2016.

Renato Assunção et al. *Computational Actuarial Science with R, V. Spatial Analysis*

Silvia Ferrari & Francisco Cribari-Neto (2004). *Beta Regression for Modelling Rates and Proportions* Journal of Applied Statistics, 31 :7, 799-815, DOI : 10.1080/0266476042000214501

A Annexe 1 : Exemple de FIRM de la ville de Des Moines, IA

Exemple : Des Moines, IA

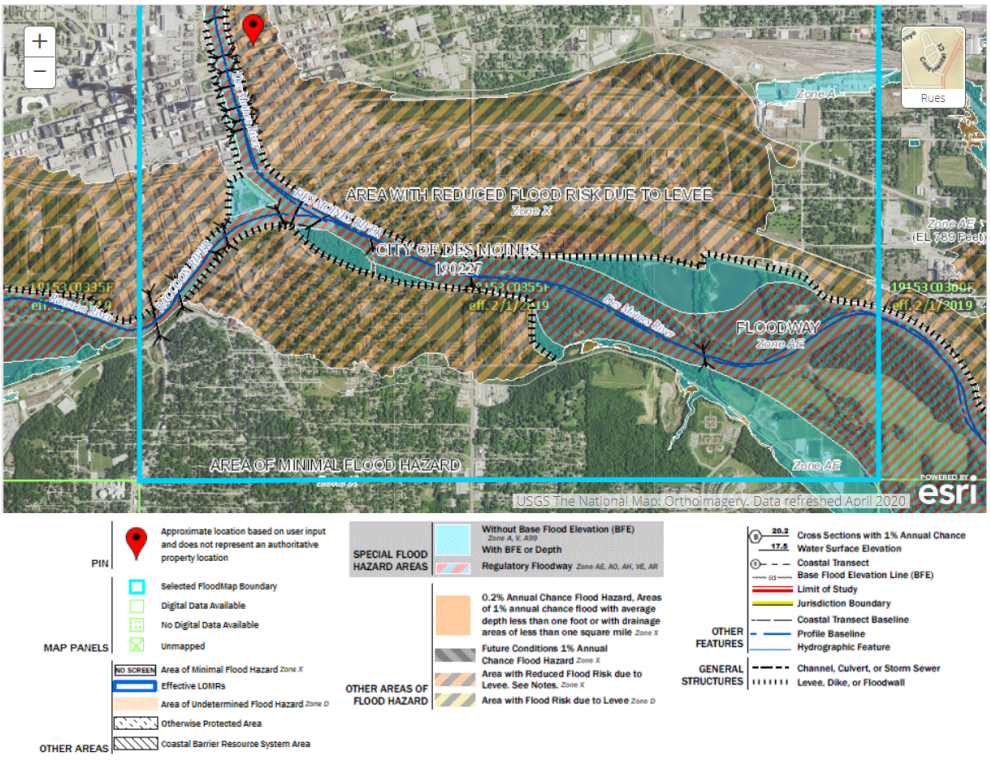


FIGURE 54: Cartographie des zones à risque de la ville de Des Moines, IA. *Flood Map Service Center*

B Annexe 2 : Etude des résultats de l'analyse spatiale avec un voisinage Queen

Indice de Moran	0.105
P-Value	7.611e-06

Table 11: Résultat du test de Moran sur les résidus de la régression linéaire classique avec un voisinage Queen

Modèle	**p-value**
LM Error	1.019e-05
RLM Error	0.5642
LM Lag	6.507e-07
RLM Lag	0.01783

Table 12: Résultats du test de Lagrange de la dépendance spatiale avec un voisinage Queen

C Annexe 3 : Corrélation des variables explicatives

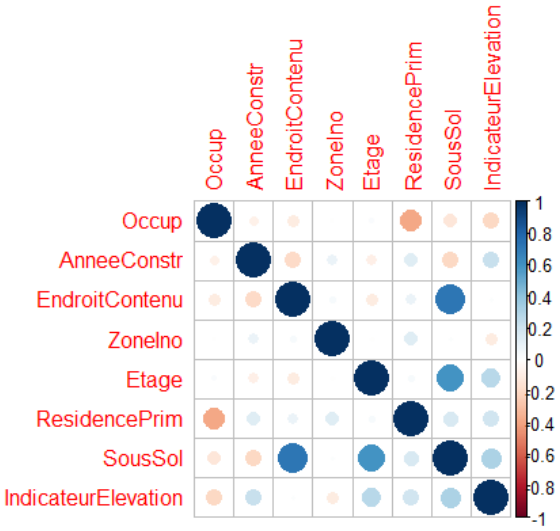


FIGURE 55: Corrélation des variables explicatives

Table des figures

Liste des tableaux