

TP: MÉTHODE DE ZONAGE PAR AGRÉGATION TERRITORIALE

L'objectif du TP est de construire un zonier en se servant notamment de la base de données `freMTPL` de la librairie `CASdatasets`. On suivra les différentes étapes décrites en cours afin de construire ce zonier.

Pour rappel, nous devons isoler les effets sur le risque dus à l'information géographique. Voici les étapes que vous devez implémenter afin de construire un zonier fictif (en effet, ici nous ne disposons pas d'une info. géographique précise telle que la commune/IRIS/...).

Etapes préliminaires :

- (1) Charger la librairie R `CASdatasets`. Importer le jeu de données `freMTPL`, plus particulièrement les sous-jeux de données `freMTPLfreq` et `freMTPLsev`.
- (2) Construire une variable qui collecte (en additionnant) l'ensemble des montants de sinistres pour les contrats multisinistrés. Concaténer cette variable au jeu de données `freMTPLfreq` afin de créer un jeu de données dans lequel une ligne représente un assuré, (multi-)sinistré ou non.
- (3) Cohérence des données : vérifier les types des facteurs de risque (covariables, ou colonnes de la base), et effectuer des changements si nécessaire. Que pensez-vous également de l'exposition ?
- (4) Grâce à la fonction `merge`, créer une base sinistrée qui pourrait être utile à la construction d'un zonier sur le coût moyen ultérieurement.
- (5) Définir un seuil de sinistralité grave à partir duquel vous créerez deux sous-bases : une base de données de sinistres attritionnels, ainsi qu'une base de données de sinistres graves.
- (6) En travaillant maintenant sur la base de fréquence, définir un individu de référence. On pourra par exemple définir chacune des modalités de référence de chaque covariable qualitative comme la modalité la plus exposée. Donner les caractéristiques de cet individu, et lister les individus de la base les ayant. Combien sont-ils ?
- (7) Créer un échantillon d'apprentissage aléatoire représentant 2/3 de la taille des données initiales. Ce tirage aléatoire pourra être stratifié ou non (en cas de stratification, stratifier sur l'exposition par exemple). Quelle est la taille de cet échantillon ? Quelle est donc la taille de l'échantillon de validation résultant ?
- (8) Quelle est la fréquence moyenne pondérée par l'exposition sur l'échantillon d'apprentissage ? La comparer à la moyenne non pondérée du nombre de sinistres sur les mêmes données. La différence est-elle importante ? Quelle est la moyenne de la fréquence annualisée ?

Modélisation de la fréquence sur la base d'apprentissage.

- (1) quelle est la moyenne empirique et la variance empirique du nombre de sinistres ? Paraît-il cohérent d'envisager une Poisson pour modéliser le nombre de sinistres ?
- (2) construire un modèle GLM log-Poisson. **On tiendra compte de l'exposition au risque sous la forme d'un offset, et on n'intégrera pas la variable Region dans la modélisation.**

- (3) Optimiser ce modèle sans perdre trop de temps. Le but est d'obtenir un modèle qui vous semble assez robuste : pour cela, vous devez vérifier sa qualité d'adéquation (résidus, tests de Wald, test LRT, comparaison empirique/modélisé sur l'échantillon d'apprentissage) et sa qualité de prévision (via la comparaison empirique/modélisé sur l'échantillon de validation).
- (4) Quelle est la moyenne du nombre de sinistres prévu sur l'échantillon de validation ? Quelle est la moyenne réellement observée ?

Construction du zonier fréquence par agrégation territoriale.

Maintenant qu'a été proprement construit le modèle de tarification sans la variable géographique, on va pouvoir travailler cet aspect puisqu'on est censé avoir isolé l'influence du facteur géographique. Toutes choses égales par ailleurs, il faut pouvoir mesurer le niveau de risque d'une "région", qui ne dépend que du facteur spatial.

- (1) On commence par calculer le risque spatial résiduel au niveau individu, tel que défini dans le cours. Le faire sur la base d'apprentissage et sur la base test.
- (2) Donner un résumé d'indicateurs statistiques simples de ce risque spatial.
- (3) On va maintenant agréger les informations des individus par région. Pour cela, on va créer un estimateur du risque spatial résiduel à l'échelle supérieure (la base de données va devenir une base dans laquelle une observation est une "région").
 - Suivre les étapes du cours pour calculer ce risque spatial à l'échelle de la région pour l'échantillon d'apprentissage.
 - Donner le risque spatial résiduel par région.
 - Créer une base de données avec pour lignes les régions étudiées et pour colonnes : le nom de la région, son exposition globale, le risque résiduel spatial correspondant, le nombre de sinistres prédits par le modèle, le nombre de sinistres observés. Vérifier la cohérence du risque spatial. Faire de même pour l'échantillon de validation, et étudier s'il y a des différences importantes.
- (4) On passe ensuite au niveau d'agrégation au-dessus : admettons que l'on connaisse la segmentation plus grossière du territoire : ILE-DE-FRANCE, HAUTE-NORMANDIE et NORD-PAS-DE-CALAIS constituent la Région1, BASSE-NORMANDIE, PAYS-DE-LA-LOIRE et BRETAGNE la Région2, POITOU-CHARENTES, AQUITAINE et LIMOUSIN la Région3, et CENTRE la Région4.
 - Créer une colonne sur la base individuelle (apprentissage et validation) du nombre de sinistres dans laquelle sera stockée l'info. de cette nouvelle région d'appartenance pour chaque individu. Quels sont les effectifs par nouvelle région ?
 - Répéter les étapes de construction du risque spatial résiduel à ce nouveau niveau (idem que la question précédente).
- (5) Reconstituer l'ensemble des informations par individu de la base de données initiale : valeur du risque spatial dans chacun des cas d'agrégation territoriale possibles. Vous créerez de nouvelles colonnes pour stocker cette information.
- (6) Il reste à définir le risque spatial résiduel conservé pour le plus petit niveau d'agrégation (hors individu!), qui sera celui-là même de la région ou celui des plus grandes régions. Il faut trouver un seuil d'exposition minimale à partir duquel la statistique du risque spatial résiduel sera considérée comme fiable. Vous procéderez comme décrit dans le cours.
- (7) Enfin, effectuer une classification par quantile d'exposition (comme décrit dans le cours) pour créer 2 classes de risque géographique. Affecter ces classes aux individus ou aux régions de plus petit niveau d'agrégation. Le zonier est terminé!
- (8) Répéter les questions (6) et ultérieures sur la base sinistrée. Le zonier construit est-il ressemblant ?