ISUP

PROMOTION 2010

Mémoire présenté devant

l'Institut de Statistique
de l'Université Pierre et Marie Curie

Pour l'obtention du

Diplôme de Statisticien

Mention Actuariat

Assurance ☒ ***Finance*** ☐

Par : M^lle^ N'ga Ahou Maïmouna COULIBALY

Sujet : Modélisation de la fréquence et du coût moyen des sinistres en Assurance Automobile du particulier avec une approche temporelle et l'intégration de variables explicatives

Lieu du stage : AXA France Solutions, Direction Technique IARD, Auto - Particuliers

Responsable du stage : Benjamin POUDRET

Invité(s) :

CONFIDENTIEL ☒

RESUME

Cette étude a été effectuée dans le cadre de l'amélioration des prévisions de sinistralité, pour le budget et les visées trimestrielles des exercices sur la branche auto. Cette branche d'AXA France a, en effet, vu sa sinistralité fortement se dégrader ; situation qui n'est pas uniquement explicable par les caractéristiques du portefeuille d'assurés. Des évènements extérieurs ont participé à cette détérioration et c'est cet aspect que nous avons voulu quantifier.

Le budget est un élément comptable dressant l'ensemble des recettes et des dépenses prévisionnelles d'un agent économique au cours d'un exercice comptable à venir. C'est un outil de prévision, d'aide au pilotage de l'entreprise. Quant aux visées, il s'agit des objectifs à atteindre. Elles sont actuellement fixées trimestriellement, mais on évolue vers des visées mensuelles pour corriger au plus vite toute dérive éventuelle.

La prévision de la sinistralité en termes de coût et de fréquence est l'un des moyens d'ajuster avec plus de précision ces éléments comptables. La dimension temporelle est un aspect que nous avons voulu conserver, pour décrire la dynamique des différentes séries de fréquences et de coûts.

Notre étude s'est basée sur le périmètre agents des véhicules à 4 roues de 2003 à 2009 en France métropolitaine. Les garanties Responsabilité Civile (RC) corporelle et matérielle, Dommages, Bris de glace, Incendie, Vol partiel et Vol total sont celles dont nous avons modélisé les fréquences et les coûts moyens mensuels. Les garanties RC et Dommages sont néanmoins les plus importantes. La charge finale prévisible en RC corporelle sur notre périmètre s'élève à 330 M€ (~27% de la charge totale) ; en RC matérielle, elle est de 290 M€ (23% du total) et en Dommages de 260M€ (21%).

La première partie de l'étude a consisté à décrire les différentes familles de variables que nous souhaitions introduire dans nos modèles :

- → les données sinistres auto 4 roues : ce sont les séries que nous voulions modéliser. C'est sur ces données que nous avons calculé les fréquences et que nous avons récupéré les séries de coûts.
- → les données portefeuille : ces données représentent les informations sur le portefeuille AXA d'assurés automobile en terme d'années police (durée d'exposition).

→ les données externes : il s'agit de diverses séries telles que des données météorologiques (la pluviométrie, les températures etc.), des données liées au véhicule (les indices des prix à la consommation sur les pièces détachées et accessoires, sur la réparation de véhicules personnels), des données liées au comportement des conducteurs (vitesses moyennes sur différents axes routiers).

Une étude descriptive de ces familles de variables a été réalisée, ainsi que l'étude des corrélations existant entre elles et nos séries cibles ; notre idée étant d'intégrer en priorité dans nos modèles les variables les plus corrélées avec la série cible, comme potentielles variables explicatives.

La seconde partie de l'étude a été consacrée à la modélisation proprement dite. Nous avons tout d'abord réalisé des modèles de type SARIMA, avec l'approche de Box et Jenkins. Ces modèles n'intègrent pas de variables explicatives. Nous nous sommes donc tournés vers une généralisation des modèles ARMA, en étudiant les modèles de régression avec erreurs ARMA. Aussi, puisque nous avions des séries qui présentaient une tendance et une saisonnalité, nous avons voulu extraire (et non supprimer) ces composantes. C'est ce qui nous a amené à faire de la modélisation des composantes inobservables (UCM), avec l'approche de Harvey et l'utilisation des modèles espace-état.

La série des fréquences mensuelles en Incendie présentait plutôt une tendance linéaire, avec des valeurs extrêmes. Pour cette série, nous avons, en parallèle de nos modélisations, effectué une régression linéaire multiple, avec le traitement des valeurs extrêmes (méthodes de régression robuste).

Après avoir ainsi obtenu nos différents modèles, une étape de validation a été nécessaire, pour tester la stabilité des modèles dans le temps. Les données ont été tronquées à fin 2008 et les prévisions des modèles sur l'année 2009 ont pu être comparées avec le réel observé sur l'année 2009.

Les modèles ayant franchi l'étape de la validation ont été comparés selon différents critères (RMSE, MAPE, MPE notamment) ainsi qu'avec les résultats du premier semestre 2010 dont nous disposions, pour nous permettre de choisir le meilleur modèle pour chaque série de fréquence et de coût.

ABSTRACT

This study is conducted within the framework of claims rate forecasting improvements, for the budget and the quarterly accounting reports on the motor line of business. This line of business in AXA France has indeed experienced a strong deterioration in its claims rate over the last few months, situation which is not only justified by the characteristics of the insured portfolio. External events participated in this deterioration and it is this aspect which we wanted to quantify.

The budget is an accounting element drawing up all the takings and the projected expenses of an economic agent during an accounting year to come. It is a tool of forecast, which helps in running the company. As for quarterly accounting reports, their aim is about objectives to reach. They are now fixed quarterly, but we evolve towards monthly reports to correct as quickly as possible any possible drift.

The forecast of claims rate in terms of cost and frequency is one of the means to adjust with more precision these accounting elements. The time dimension is an aspect which we wanted to keep, to describe the dynamics of the various frequencies and costs series.

Our study bases itself on the 4 wheels tied agents perimeter from 2003 till 2009 in Metropolitan France. Third party liability both physical and material, Damage, Glass breakage, Fire, partial Theft and total Theft are the covers whose frequencies and monthly average costs were modeled. The first three covers mentioned above are nevertheless the most important. The expected final load in third party liability with bodily injuries on our perimeter amounts to 330 M€ (~27 % of the total load); in material damage to a third party it is 290 M€ (~23 % of the total) and in Damage it is 260M€ (~21 %).

The first part of the study consists in describing the various families of variables which we wish to introduce into our models:

- → The 4 wheels claims data: they are the series which we want to model. It is on these data that we compute the frequencies and those on which we get back series of costs.
- → The portfolio data: these data represent the information on the AXA France private car insured portfolio in terms of years policy (exposure time).

→ The external data: it is about diverse series such as meteorological data (rainfall, temperatures etc.), data connected to the vehicle (consumer price index on the spare and secondary parts, on the repair of personal vehicles), data connected to the behavior of drivers (average speeds on various main highways).

A descriptive study of these families of variables is realized, as well as the study of the correlations existing between them and our target series; our idea being to integrate first and foremost into our models the most correlated variables to the target series, as potential explanatory variables.

The second part of the study is dedicated to the modeling itself. At first, we carries out SARIMA's type models according to Box and Jenkins' approach. These models do not integrate explanatory variables. We thus turn to a generalization of the ARIMA models, by studying the models of regression with ARMA errors. So, because we have series which present a trend and/or seasonality, we want to extract (and not delete) these components. This is why we process unobservable components models (UCM), with the approach of Harvey and the use of the space-state models. The monthly frequencies series in Fire coverage presents rather a linear trend, with extreme values. For this series, we make, in parallel of our modeling, a multiple linear regression, with the treatment of extreme values (methods of strong regression).

Having thus obtained our various models, a stage of validation is necessary to test the consistency of the models over time. The data have been truncated at the end of 2008 and the forecasts of the models over the year 2009 enable us to compare these with the real ones observed over the year 2009.

The models having succeeded the stage of the validation are finally compared according to various criteria (RMSE, MAPE, MPE in particular) as well as with the results of the first half of the year 2010, to allow us to choose the best model for every series of frequency and cost.

REMERCIEMENTS

Je tiens à remercier Monsieur Benjamin POUDRET, mon maître de stage, qui m'a encadrée et accueillie au sein de son équipe. Il est d'ailleurs l'initiateur de cette approche de la modélisation des fréquences et coûts. Je lui dois un grand merci pour les multiples relectures, les conseils et le temps qu'il a consacré à m'expliquer la démarche et la méthodologie à adopter.

Je tiens également à remercier tous mes collègues du service auto-particuliers au sein de la Direction Technique IARD de AXA France, grâce auxquels chaque journée de travail était ponctuée de moments de détente et avec lesquels j'ai pu échanger, apprendre et travailler durant ces six mois. Je ne pouvais imaginer une meilleure intégration que celle effectuée dans ce service.

Je remercie de même
Monsieur Olivier LOPEZ, mon responsable pédagogique à l'ISUP, pour avoir pris le temps d'écouter mes nombreuses doléances et pour les corrections qu'il a apporté à cette étude ;
Madame Marie KRATZ, professeur de séries temporelles à l'ISUP, pour m'avoir donné envie d'explorer ce sujet et pour toutes les références qu'elle m'a fournie.

A tous ceux et celles qui, de près ou de loin, ont suivi avec intérêt ce sujet de mémoire, j'adresse aussi mes remerciements.

Je ne pourrai terminer sans remercier ceux qui m'ont toujours soutenu dans mon parcours scolaire, mes parents, mes sœurs, et ma tante ; ce sont les personnes qui ont cru en moi et pour lesquelles j'ai toujours voulu donner le meilleur de moi dans mon travail.

SOMMAIRE

INTRODUCTION

Les assurances dommages incluent les assurances de biens et les assurances de responsabilité. Très affecté par les retombées de la crise économique et une concurrence accrue, ce secteur a connu en 2009 une progression très modeste (+1 %) et inférieure à celle observée en 2008 (+2,5 %)[1].

Les assureurs français se sont résolus à augmenter ou du moins à stabiliser les tarifs des assurances dommages depuis le 1er janvier 2009. Cette tendance se poursuit en 2010. Auparavant, la forte concurrence dans ce domaine avait entraîné une importante baisse des primes. Ce qui a eu pour effet, entre autres, de fragiliser, voire même dégrader les résultats techniques de la branche automobile.

En 2008 les sinistres matériels avaient chuté de 4,9 % et les sinistres corporels de 5 %. En 2009, leur nombre a progressé :

+ 2 % pour les accidents en « Responsabilité Civile » ;
+ 9 % pour les accidents en « Dommages ».

Les mauvaises conditions climatiques qui ont touché la France en 2009 sont une autre des raisons de la dégradation de la sinistralité (tempête Klaus en janvier, tempête Quinten en février, orages de grêle en mai, un épisode de grand froid). L'année 2010 annonce déjà des sinistres importants à payer, avec la tempête Xynthia d'avril.

Pour le Groupe AXA, l'assurance dommages représente 6,7 Milliards d'Euros [2](soit 26,2 %) de son chiffre d'affaires global. Les défis dans ce secteur pour les années à venir sont donc importants du fait de l'accroissement du montant des sinistres de dommages, de l'apparition de nouveaux sinistres de responsabilité et des incertitudes croissantes sur les montants de sinistres maximum possibles, rendus plus difficiles à évaluer et à prévoir sur les bases historiques actuelles.

Chaque fin d'année, l'actuariat Auto rédige un budget avec plusieurs hypothèses : effet prix (variation de la prime moyenne du portefeuille sur un an), volumétrie d'affaires nouvelles etc.

[1] Etudes et Statistiques du 27 janvier 2010 – Fédération Française des Sociétés d'Assurance
[2] Au 31 décembre 2009 - Source : FFSA – Périmètre d'affaires directes des groupes hors IP et Mutuelles Santé

Pour réajuster certaines hypothèses pas forcément vérifiées (apport net, tempêtes plus ou moins violentes, …), on réalise des visées trimestrielles. C'est dans ce cadre de maîtrise de plus en plus indispensable de la sinistralité que s'inscrit cette étude sur la modélisation de la fréquence et du coût moyen des sinistres en assurance automobile des particuliers.

Pour modéliser les fréquences nous analyserons plusieurs types de modèles, en privilégiant la prise en compte de l'effet temporel. C'est-à-dire que nous modéliserons la fréquence comme une série temporelle, une réalisation d'une famille de variables aléatoires indicées par le temps. Pour chaque garantie, nous étudierons la série de fréquences mensuelles avec une recherche de saisonnalité, de tendance, et l'intégration de variables exogènes susceptibles d'améliorer le modèle. Pour la modélisation des coûts, nous tenterons également une modélisation de type série temporelle.

Après validation, une comparaison des différents modèles obtenus sera faite avec des critères tels que le MAPE ou le RMSE pour pouvoir choisir le meilleur modèle pour chaque série de fréquences et de coûts.

Les analyses se feront à un niveau macroscopique, c'est-à-dire que la vision que nous aurons sera à la maille portefeuille. Nous ne descendrons pas à la maille client.

I. CONTEXTE DE L'ETUDE

I.1. Enjeux et problématique

Dans son article du 16 octobre 2009 sur l'assurance automobile, la FFSA[3] a constaté une augmentation du coût global de la charge de sinistres liée à des accidents automobiles. Très concrètement, une augmentation de 2 % des accidents responsabilité civile matériels et de 9 % des accidents dommages a été enregistrée. Cette dégradation semblait se confirmer pour le second semestre 2009 au vu des chiffres publiés par la sécurité routière, qui font état d'une augmentation de 8,5 % des accidents recensés.

Aussi, l'augmentation du coût moyen des sinistres corporels se poursuit. Elle est particulièrement sensible sur les sinistres corporels "lourds" : si le nombre de blessés et de tués sur la route a baissé de 40 % de 1997 à 2007, le nombre de sinistres corporels supérieurs à 3 millions d'euros enregistrés par les assureurs, a été multiplié par près de six sur la même période.

Sur la période 2002-2008, la nette amélioration de la mortalité routière est due essentiellement au renforcement des contrôles routiers qui a eu un effet à la baisse sur les vitesses moyennes pratiquées sur l'ensemble des réseaux (–1,4 % par an en moyenne sur la même période). Sur cette même période, les sinistres automobiles ont enregistré une baisse sensible de leur nombre (près de 3 % par an en moyenne)[4], qu'il s'agisse d'accidents purement matériels, d'accidents à l'origine de dommages corporels, de vols de véhicules ou de dommages et de bris de glace.

A contrario, l'exercice 2009 se caractérise par une reprise à la hausse des fréquences, une augmentation de la circulation et une légère remontée des vitesses pratiquées. Le contexte climatique particulièrement défavorable de cette année est venu renforcer la charge des garanties dommages et bris de glace (tempêtes Klaus, Quinten, …). Notons que même si cet élément est conjoncturel, il est fort probable que la fréquence de ces sinistres s'accentue du fait du réchauffement climatique.

[3] Fédération Française des Sociétés d'Assurance
[4] Revue Risques - ÉVOLUTION RÉCENTE DE L'ASSURANCE DE BIENS DES PARTICULIERS J. Cornu, B. Gatterer

Il ressort de ce bref état des lieux que la sinistralité se trouve influencée par des paramètres qui ne sont pas forcément liés aux caractéristiques des assurés. La modification de l'un ou l'autre de ces paramètres pourrait expliquer, dans une certaine mesure, la sinistralité observée.

La politique AXA pour le pilotage de l'activité est d'établir un budget. Le budget d'un exercice est la prévision du bilan comptable avant le début de l'exercice concerné. Il prévoit les charges ainsi que les recettes de la compagnie d'assurance. C'est un outil dynamique de gestion, dont la mise en place nécessite différentes hypothèses d'évolution du chiffre d'affaires et de la sinistralité (actuellement formulées en se basant sur les chiffres d'affaires passés et les dernières tendances actuelles).

Lors de l'adoption du budget, il est impératif de pouvoir prédire la baisse ou la hausse de la sinistralité. L'écart entre la sinistralité « constatée » en fin d'année et les hypothèses prises en début d'année doit être réduit au maximum. L'enjeu n'est pas seulement de retarifer, mais aussi de pouvoir expliquer et prévoir les éventuelles dérives, par des actions correctives lors des visées. Ces dernières sont des révisions du budget à la fin du premier, deuxième et troisième trimestre de l'exercice.

Un simple regard sur la sinistralité passée n'est pas suffisant pour nos hypothèses d'évolution; cette évolution doit pouvoir être expliquée au mieux.

Les objectifs de cette étude seront de déterminer et de modéliser l'influence des valeurs passées et des facteurs potentiels, qu'ils soient internes ou externes, qui impactent la sinistralité de notre portefeuille, en termes de coût et de fréquence.

I.2. Environnement de travail

Le groupe AXA est aujourd'hui l'un des leaders mondiaux de la Protection Financière. Présent dans 57 pays, avec 96 millions de clients, les activités du groupe sont géographiquement diversifiées, avec une concentration sur les marchés d'Europe, d'Amérique du Nord et de la région Asie-Pacifique.

En France, AXA France propose à ses clients une gamme complète de produits et services en assurance vie, assurance dommages, banque, assistance, protection juridique, santé et prévoyance, épargne, transmission de patrimoine.

Elle est divisée en plusieurs entités organisées comme suit :

Tableau 1 : Les entités de AXA France

L'étude qui suit a été réalisée au sein d'AXA France Solutions, à la Direction Technique IARD, dans le service Risques de Masse, au pôle Auto – Particuliers.

Le pôle Auto – Particuliers a comme mission de créer et de tarifer des produits d'assurance automobile, pour des véhicules à 4 roues ou 2 roues, destinés à la clientèle « particuliers » pour un usage privé et/ou professionnel. Il s'occupe aussi du suivi de ces produits en fournissant les indicateurs utiles aux autres services et entités.
Il comprend l'équipe Actuariat (au sein de laquelle cette étude a été effectuée) et l'équipe Produits.

AXA France dispose de trois réseaux de distribution :

- un réseau d'agents AXA : ils sont mandatés par l'assureur et reçoivent une commission pour chaque contrat vendu ;
- un réseau de courtiers : mandatés par le client, ils sont chargés de trouver les meilleurs prix sur le marché ;
- un réseau de salariés AXA : ils font de la vente à domicile.

Les réseaux de distribution n'ont pas les mêmes comportements. Mais seul le périmètre Agents sera étudié, car étant le plus important des trois (85 % des contrats en portefeuille). La période d'étude couvrira les années 2003 à 2009.

I.3. Les principales garanties

En assurance automobile des particuliers, il y a quatre principales garanties :

- La garantie Responsabilité Civile
- La garantie Dommages accidentels
- La garantie Incendie-Vol
- La garantie Bris de glace

La seule garantie obligatoire en automobile en France est la Responsabilité Civile. Toutes les autres sont facultatives.

Dans l'offre AXA, ces garanties sont déclinées en trois niveaux auxquelles s'ajoutent des options complémentaires à la carte et des packs.

	Garanties de base	Options à la carte	Packs
Niveau 1	• Responsabilité civile • Défense pénale et recours suite à accident (D.P.R.S.A.) • Protection juridique (10 000 €) • Sécurité du conducteur (450 000 €) ***** • Décès du conducteur (10 000 €) • Assistance aux personnes (franchise 30 km) • Capital réparation (1 500 €) • **Sans antécédents :** Joker (– 25 ans) • **Assurance Auto des Pros** : Avantage Privé-pro	• Assistance accident et panne (sans franchise kilométrique) • Sécurité du conducteur étendue (1 000 000 € - franchise A.I.P.P. 10 %) • Bris des glaces sans franchise • Atout' Âge • Kilométrage limité (– 8 000 km / – 10 000 km pour les pros) • Sans antécédents : stage de conduite • Forfait 8000 : conduite exclusive et franchises proportionnelles	•**Pack « Ma Sécurité »** - SDC étendue (1 000 000 € franchise **A.I.P.P. 5 %)** - **GAV** (1 000 000 € franchise **A.I.P.P. 5 %)** • **Pack « Ma Mobilité »** - Assistance au véhicule - Véhicule de remplacement suite à panne (7 j), accident (15 j), vol (30 j)

Niveau 2	**Niveau 1 +** • Incendie-Vol • Bris des glaces sans franchise • Attentats • Événements climatiques • Catastrophes naturelles • Catastrophes technologiques • Aménagements et accessoires (3 500 €) • Formules Monospace : Effets personnels, autoradio/GPS (1 000 €) • Assurance Auto des Professionnels : Effets personnels, professionnels, autoradio/GPS (1 000 €)	**Options Niveau 1 +** • Accessoires et aménagements > 3 500 € (dans la limite de 20 000 €) • Effets personnels, autoradio/GPS (1 000 €, 1 500 €, 3 000 €) • Majoration, rachat partiel ou total de la franchise Incendie-Vol	
Niveau 3	**Niveau 2 +** • Dommages tous accidents • Valeur à neuf 12 mois	**Options Niveau 2 +** • Valeur à neuf 24, 36 ou 60 mois, suivie de la VADE + 15 % • Majoration, rachat partiel de la franchise Dommages tous accidents	**• Pack niveaux 1 - 2** **ET / OU** • Pack « Mes + Auto » - Valeur à neuf 24, 36 ou 60 mois, suivie de la VADE + 15 % - Accessoires et aménagements > 3 500 € (dans la limite de 20 000 €) **OU** **• Pack « Excellence »** - Toutes les garanties des packs « Ma Mobilité » et « Mes + Auto » + - Sécurité du conducteur étendue (1 000 000 € – franchise **A.I.P.P. 10 %**) - Assistance enrichie - Véhicule de remplacement à l'identique - Rachat total de la franchise en Incendie-Vol - Rachat partiel de la franchise Dommages tous accidents, abrogation de la franchise Dommages tous accidents **en cas de sinistre survenu avec un tiers identifié**

*** Si l'assuré a souscrit un contrat Garantie Personnelle du Conducteur, la Garantie Sécurité du Conducteur peut être exclue. :**

Tableau 2 : Récapitulatif des niveaux de garanties Auto AXA

a. La garantie Responsabilité Civile

Elle permet l'indemnisation des dommages causés aux tiers par le conducteur du véhicule ou un passager. Elle couvre uniquement les dommages causés aux tiers donc exclut le véhicule assuré, ses passagers ainsi que son conducteur.

Les dommages corporels causés aux tiers dans le cadre d'un accident responsable seront pris en charge au titre de la garantie Responsabilité Civile Corporelle. Dans ce cas, le montant de la garantie est illimité.

Les dommages matériels causés aux tiers seront eux enregistrés au titre de la garantie Responsabilité Civile Matérielle. Le montant de la garantie est ici plafonné à 100 millions d'euros ; mais si la garde ou la conduite du véhicule a été obtenue contre le gré du conducteur, le plafond baisse à 460 000 €.

b. La garantie Dommages tous accidents

Cette garantie permet de couvrir les dommages matériels causés en l'absence de tiers responsable, lors d'une collision avec un ou plusieurs véhicules, d'un choc avec un corps fixe ou mobile, d'un versement sans collision préalable ou d'un acte de vandalisme.

c. La garantie Incendie

Elle permet de recevoir une indemnisation pour les dommages subis par le véhicule assuré causés par un incendie, l'action de la foudre ou une explosion (à l'exception de l'explosion des pneumatiques et des dommages en résultant).

d. La garantie Vol

La garantie Vol permet de couvrir les dommages résultant de la disparition du véhicule assuré ou de sa détérioration à la suite d'un vol ou d'une tentative de vol. Elle garantit aussi les dommages résultant de la disparition ou des éléments volés indépendamment du véhicule, s'ils entrent dans la définition du véhicule assuré.

Elle est subdivisée en 2 catégories :

- → La garantie Vol total : cas où le véhicule n'est pas retrouvé.
- → La garantie Vol partiel : cas où le véhicule est retrouvé.

Les garanties Incendie et Vol sont souscrites en même temps et constituent la garantie Incendie-Vol.

e. La garantie Bris de glace

La garantie Bris de Glace couvre les dommages subis par le pare-brise, la vitre arrière, l'ensemble des feux avants y compris les ampoules et le toit ouvrant ou non du véhicule assuré. Par contre les feux arrières, les rétroviseurs ou tout autre élément en verre, glace, ou verre organique ne sont pas couverts par cette garantie.

La répartition des sinistres des assurés d'AXA France selon les différentes unités de prestation (en abrégé UP) est la suivante pour l'année 2009 :

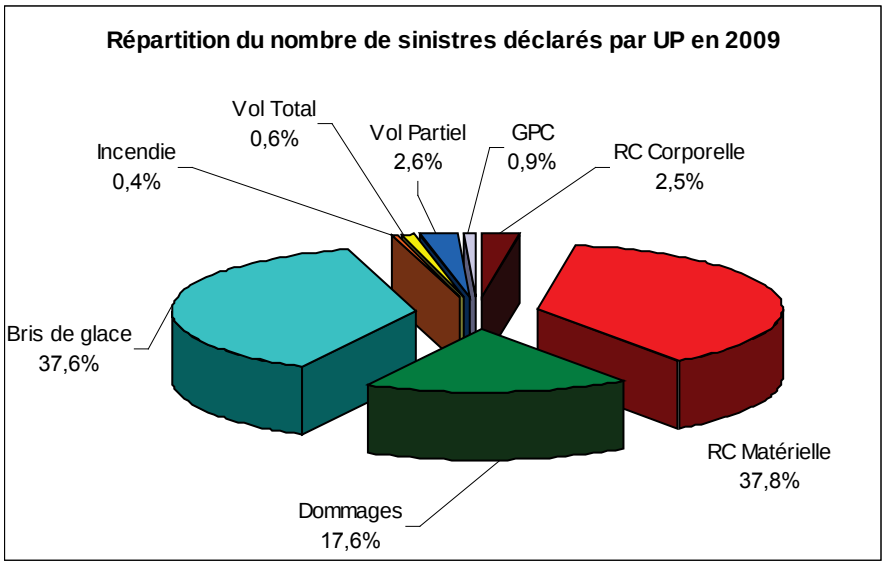


Figure 1 : Répartition du nombre de sinistres par unité de prestation

Les garanties représentées ci-dessus, excepté la GPC (Garantie pour le conducteur) sont celles qui seront étudiées dans ce mémoire. En effet, la souscription de la GPC ne s'est accrue que depuis 2007 et nous ne disposions pas d'un historique suffisant pour pouvoir effectuer une étude sur cette garantie.

II. PRESENTATION DES DONNEES

II.1. Présentation des variables

Avant de passer à la présentation des variables, il est important de comprendre comment elles ont été enregistrées ainsi que le fonctionnement des bases dont elles ont été extraites. Deux types de bases internes ont été nécessaires pour récupérer nos données. Il s'agit de la base du Statauto et des bases sinistres.

La base du Statauto : Le Statauto est la base de données automobile de la Direction Technique Auto Particuliers d'AXA. Elle contient les données relatives à l'auto, la moto ou au cyclomoteur. Cette base est constituée d'images à l'instant t de chaque contrat. Une image est une photo du contrat à un instant donné. Il s'agit d'une portion de la police pendant laquelle aucune modification des conditions d'assurances n'est survenue.

Dans le Statauto, une ligne correspond à une image et chaque image contient de nombreuses caractéristiques du contrat, de l'assuré et de son véhicule.

Dès qu'il survient un fait de production (affaire nouvelle, remplacement, remise en vigueur) ou un terme (échéance principale) sur un contrat, une image supplémentaire est générée donc une ligne supplémentaire. Toutes les images des contrats de l'année sont conservées. A chaque image est associée une nature de début d'image qui correspond à l'état du contrat au début de la période couverte par l'image et une nature de fin d'image.

Les natures de début et fin d'image possibles sont les suivantes :

Natures de début d'image	***Natures de fin d'image***
une affaire nouvelle	une résiliation
un avenant / remplacement	un avenant / remplacement
une remise en vigueur	une suspension
une échéance principale	une échéance principale
un sans effet	

Parmi les informations du Statauto, nous avons utilisé les RT (Résultats Techniques) : ce sont les images de nos contrats sur 12 mois glissants. Chacune des images a les différentes caractéristiques de chacun des contrats.
Dans notre base, pour chaque année d'étude N, toutes les images ont été sélectionnées sur la période du 1er janvier N au 31 décembre N, vue à fin février N+1. C'est à partir de cette base que nous avons obtenu les données portefeuille que nous détaillerons dans la suite.

On appelle année police ou exposition au risque, la durée de l'image contrat. L'exposition au risque étant très variable selon l'image contrat, pour toutes les procédures statistiques employées sur les données portefeuille, la pondération par les années police a été utilisée.

Les bases sinistres de l'INFOCENTRE extraites des bases ACN:

Ces bases contiennent deux types d'informations sinistres :

- un système de tables contenant les informations des Sinistres selon leur date de survenance quel que soit leur état ou leur date de clôture ;
- un système de tables contenant les informations des Sinistres 'En cours ou clos depuis moins de 2 ans', quelle que soit leur année de survenance.

Les tables par année de sinistre contiennent trois types de librairies : la librairie concernant les sinistres Auto, la librairie concernant les sinistres non-Auto et la librairie concernant les règlements. Les tables des « sinistres en cours ou clos depuis moins de 2 ans » contiennent quatre librairies : une première sur les sinistres Auto, une deuxième sur les sinistres non-Auto, une troisième sur les sinistres Construction et une quatrième sur les règlements.
Dans ces différentes tables, nous ne sommes intéressés qu'aux sinistres Auto.

Dans les bases ACN, les données sinistres sont présentées de manière plus détaillée que dans le Statauto, pour chaque type de garantie. C'est de ces bases que nous avons extrait toutes les informations relatives aux nombres et aux coûts des sinistres suivant le type de garantie. Les coûts sont exprimés en euros et représentent les coûts moyens mensuels.

En RC corporelle, les coûts de sinistres étaient écrêtés à 90 000 € jusqu'en 2004 puis à partir d'avril 2004 l'écrêtement est passé à 150 000 €.

Les variables utilisées ont été regroupées en famille décrites ci-dessous (du II.1.1 au II.1.3.).

II.1.1. Les données sinistres Auto 4 roues

Nous disposons, pour chacune des garanties de notre étude, des variables suivantes :

- Le nombre de sinistres déclarés (mensuels, cumuls janvier, et 12 mois glissants) au cours de l'année, qu'ils soient clos ou non
- Le coût moyen des sinistres déclarés (qu'ils soient clos ou non) avec une survenance dans l'année courante
- Le coût moyen de tous les sinistres clos dans l'année, quelles que soient leurs dates de survenance.

Ces données sont mensuelles de janvier 2003 à décembre 2009, et ne portent que sur le réseau des agents en Auto 4 Roues. Elles nous sont fournies par le pôle Actuariat et Réassurance de la Direction technique IARD.

II.1.2. Les données portefeuille

Les séries de données concernant le portefeuille d'assurés ont été extraites des fichiers RT contenus dans le Statauto. Les variables que nous avons retenues ont été classées par famille et sont présentées ci-dessous :

Variables liées au conducteur :

Le R/M : c'est le coefficient de réduction/majoration qui s'applique sur la prime de référence. Il s'agit d'un système de bonification et de pénalisation pour les conducteurs, communément appelé système du « Bonus – Malus ». Il se calcule sur une échelle hiérarchique de 50 à 350. Nous avons effectué un regroupement des modalités en conservant la classe des bonus 50, celle des bonus 100 (tout nouvel assuré passe en bonus 100), la classe des bonus de plus de 100 (ce sont des malus). La méthode utilisée pour la description des classes est détaillée dans l'annexe 3.

L'ancienneté de Bonus 50 et le RM ont été croisées, car les assurés Bonus 50 représentent 60 % du portefeuille. La variable ancienneté de Bonus 50 nous aide à segmenter la classe, selon la durée de détention du Bonus 50. Elle a comme modalités le nombre d'années selon l'ancienneté de Bonus 50 ou le non Bonus 50 du conducteur.

La variable Ancienneté de permis : Cette variable a été reconstruite en utilisant la variable « nombre de mois écoulés depuis l'obtention du permis » disponible dans le Statauto.

L'ensemble des variables de ce groupe est résumé dans le tableau ci-dessous :

Variable	Modalité
Age du conducteur	• 18 – 25 ans • 26 – 35 ans • 36 – 45 ans • 46 – 55 ans • 56 – 70 ans • + de 70 ans • personne morale
Sexe	1 : homme 2 : femme 3 : personne morale
Situation matrimoniale	1 : célibataire 2 : concubin 3 : marié 4 : séparé 5 : divorcé 6 : veuf
Antécédents à la souscription	1 : sans antécédents 2 : avec antécédents
Usage du véhicule	1 : privé 2 : privé-trajet 3 : professionnel 4 : tournées 5 : non PE044
Nombre de véhicules du foyer	1 : 1 véhicule 2 : 2 véhicules 3 : 3 véhicules 4 : 4 véhicules 5 : 5 véhicules et plus
Nombre de permis du foyer	1 : 1 permis 2 : 2 permis 3 : 3 permis 4 : 4 permis 5 : 5 permis et plus
Ancienneté de Bonus 50 * RM	• Bonus 50 depuis moins de 3 ans • Bonus 50 depuis 3 à 6 ans • Bonus 50 depuis 6 ans et plus • Bonus 51 – 60 • Bonus 61 – 74 • Bonus 75 – 99 • Bonus 100 • Bonus supérieur à 100
Ancienneté de permis	• Moins de 2 ans • 3 – 6 ans • 7 – 10 ans • 11 – 14 ans • 15 – 20 ans • 21 – 30 ans • 31 ans et plus

Variables liées au contrat et aux clauses :

La variable Régions AXA France : C'est une variable différente des régions françaises légales. Elle correspond aux 5 grandes régions AXA : Ile de France, Ouest, Nord-est, Sud-ouest, Sud-est, auxquelles s'ajoutent 4 autres classes : les DOM-TOM notés EOM, la Mutuelle Saint-Christophe notée MSC (assurance pour le clergé), le groupe NATIO (portefeuille en run-off détenu avec BNP Paribas), et AXA Partenaires (assurance des salariés d'AXA).

Figure 2 : Carte des régions AXA France

Les variables Zone Vol et Zone RC : Le zonage est un regroupement des communes françaises suivant leur sinistralité. Le classement est effectué des communes les moins exposées aux sinistres, dans la première zone, aux communes les plus exposées dans la dernière. Les zones sont définies par garanties et non par formule. Ce qui veut dire qu'une classification est effectuée pour le Vol et une autre, distincte, pour la RC.

La variable Kilométrage limité : La promotion de forfaits automobiles à kilométrage limité, ou de réductions de tarifs graduelles en fonction des distances parcourues, a été mise à l'étude par certains assureurs depuis de nombreuses années, et ce afin de réduire indirectement le taux de sinistralité (en incitant les assurés à moins utiliser leur véhicule).

La variable Statut juridique : Il s'agit du statut juridique de la société AXA à laquelle est rattaché le portefeuille de l'agent chez qui le contrat est souscrit.

L'ensemble des variables de ce groupe est résumé dans le tableau ci-dessous :

Variable	Modalité
Ancienneté du contrat	1 : moins d'un an 2 : 1 – 2 ans 3 : 2 – 3 ans 4 : 3 – 4 ans 5 : 4 – 5 ans 6 : 5 – 6 ans 7 : 6 – 8 ans 8 : 9 – 13 ans 9 : 14 ans et plus
Régions AXA France	1 : Ile de France 2 : Nord-est 3 : Sud-est (hors Michelin) 4 : Sud-ouest 5 : Ouest 6 : EOM 7 : MSC 8 : Natio 9 : AXA Partenaires
Zone VOL	0 : EOM (Dom Tom) 1 : Zone A 2 : Zone B 3 : Zone C 4 : Zone D 5 : Zone E 6 : Zone F 7 : Zone G 8 : Zone H 9 : Zone I 10 : Zone J 11 : Zone K 12 : Zone L 13 : Zone M 14 : Zone N
Zone RC	0 : EOM 1 : Zone 1 2 : Zone 2 3 : Zone 3 4 : Zone 4 5 : Zone 5 6 : Zone 6 7 : Zone 7 8 : Zone 8 9 : Zone 9 10 : Zone 10 11 : Zone 11 12 : Zone 12 13 : Zone 13
Garantie Dommages	1 : franchise majorée 2 : franchise normale 3 : rachat total 4 : rachat partiel 5 : collision pdts migr.10, 20, 30% 6 : proportionnelle (option 8000K) 7 : autres cas 8 : non souscrite ou indéterminée
Garantie Bris de glace	1 : avec franchise 2 : sans franchise 3 : franchise majorée 4 : proportionnelle

	5 : non souscrite ou indéterminée
Garantie Incendie-Vol	1 : franchise majorée 2 : franchise de base 3 : rachat total 4 : rachat partiel 5 : proportionnelle (option 8000K) 6 : souscrite autres cas 7 : non souscrite
Garantie du conducteur	1 : Garantie personnelle du conducteur 2 : Sécurité du Conducteur de base 3 : Sécurité du Conducteur étendue 4 : non souscrite
Kilométrage limité	1 : souscrite 8000 (y.c. Forfait 8000) 2 : non souscrite 3 : 9000 kms essence (UAP) 4 : 9000 kms diesel (UAP) 5 : 12000 kms diesel (UAP) 6 : 10000 kms (TNS usage pro) 7 : forfait 4000 (clause RPY)
Statut juridique	1 : Société anonyme 2 : Mutuelle

Variables liées au véhicule :

La variable Age du véhicule : Cette variable a été construite à partir des renseignements sur l'année et le mois de première mise en circulation du véhicule.

La variable Classe de prix :

Sur l'ensemble de la période d'étude, nous avons ensuite effectué un regroupement des classes SRA. Pour harmoniser les classes de prix des véhicules sur la période d'étude, il a fallu transcrire en classes de prix SRA[5] les classes de prix APSAD[6] pour les années 2003 et 2004 (annexe 3).

Variable	Modalité
Code énergie (type de carburant utilisé par le véhicule)	1 : Essence 2 : Diesel 3 : Electrique 4 : G.P.L. 5 : Gaz naturel 6 : Bioéthanol
Alimentation du véhicule (type de système d'alimentation utilisé par le véhicule)	1 : Electrique 2 : GPL 3 : Carburateur 4 : Injection directe suralimentée 5 : Injection directe 6 : Injection directe suralimenté 7 : Gaz naturelle de ville 8 : Hydrogène-essence

[5] Sécurité Réparation Automobiles
[6] Assemblée Plénière des Sociétés d'Assurance Dommages

Vitesse maximale du véhicule	1 : <=140 2 : 141-150 3 : 151-160 4 : 161-170 5 : 171-180 6 : 181-190 7 : 191-200 8 : 201-220 9 : > 220
Age du véhicule	• Moins d'1an • 1-2 ans • 3-4 ans • 5-6 ans • 7-8 ans • 9-10 ans • 11-14 ans • +de 15 ans
Ancienneté de la carte grise	1 : moins d'1an 2 : 1-2 ans 3 : 2-3 ans 4 : 3-4 ans 5 : 4-5 ans 6 : 5-6 ans 7 : 6-8 ans 8 : 8-9 ans 9 : 10 ans et +
Classe de prix	1 : 0 - 10 500€ 2 : 10 500€ - 16 000€ 3 : 16 000€ - 20 150€ 4 : 20 150€ - 25 500€ 5 : 25 500€ - 32 000€ 6 : 32 000€ - 85 500€ 7 : 85 500€ - 1 524 490€

Pour chacune des variables du portefeuille, nous avons raisonné en termes d'années police. Pour chaque modalité, c'est la somme des années police par année que nous avons utilisé comme donnée. Pour chaque variable, l'information que nous avons retenue est la proportion d'assurés dans le portefeuille qui possède telle ou telle modalité.

II.1.3 Les données externes

Elles ont été classées par type :

- Les variables météorologiques :
  - La pluviométrie : exprimée en millimètres
  - L'ensoleillement : exprimé en heures de durée d'insolation
  - Les températures : exprimées en degrés Celsius.
  - Le gel : exprimé en nombre de jours avec gel c'est-à-dire les journées pendant lesquelles la température est inférieure ou égale à 0° C.

Ces données moyennes mensuelles nous ont été fournies par Météo France, de janvier 1979 à décembre 2009.

- Les variables relatives au véhicule :
  - L'IPC (indice des prix à la consommation) pour les carburants et les lubrifiants
  - L'IPC pour les pièces détachées et accessoires de véhicules personnels
  - L'IPC pour l'entretien et la réparation de véhicules personnels

  Ces indices fournis par l'INSEE (Institut National de la Statistiques et des Etudes Economiques) sont mensuels, de janvier 1998 à décembre 2009, en base 100. L'indice des prix à la consommation permet de « mesurer » l'inflation. Il nous donne l'évolution des prix à la consommation par rapport à l'année de référence qui est ici 1998.

  - Les indices de parcours mensuels sur différents axes routes : réseau non concédé total, autoroutes interurbaines, autoroutes et voies rapides urbaines, routes nationales interurbaines à caractéristiques autoroutières, autres types de routes nationales, réseau national (moyenne mensuelle), autoroutes non concédées (moyenne mensuelle), total des autoroutes (moyenne mensuelle), total des routes nationales (moyenne mensuelle).

  Ils nous sont fournis par le Sétra (Service d'études sur les transports, les routes, et leurs aménagements) de janvier 2001 à décembre 2009, avec comme unité le milliard de véhicules par kilomètre.

  - Le volume mensuel de livraison de carburants (Super et Gazole) en m^3.
  - Les prix des carburants (Super sans plomb 95, Super sans plomb 99, GPL, Gazole) en € TTC par litre.

  Ces données sont mensuelles, disponibles de janvier 2004 à décembre 2009. Elles sont fournies par le CPDP (Comité Professionnel Du Pétrole).

  - Le nombre mensuel d'immatriculations de véhicules particuliers (neufs et d'occasion) en France

  Ces données sont fournies par le SOeS (Service de l'Observation et des Statistiques) et le FCA (Fichier Central des Automobiles), de 2002 à 2009.

- Variables liées aux comportements des assurés :
  - Le kilométrage moyen total des ménages

    Ces données sont annuelles de 1980 à 2008, fournies par la SOFRES (Société Française d'Enquêtes par Sondages).

  - Les vitesses moyennes de jour et de nuit sur les différents axes routiers (autoroutes, RN, RD, Artères en agglomération, Entrée et Sortie des agglomérations) en unité

    Ces données sont quadrimestrielles, de 1996 à 2009, fournies par l'ONISR (Observatoire National Interministériel de la Sécurité Routière) et la DREIF (Direction Régionale de l'Equipement d'Ile de France).

  - Le nombre d'accidents corporels et de victimes (blessées et tuées), de 1980 à 2009.

    Ces données sont mensuelles, de 1980 à 2009, et proviennent de l'ONISR.

Notre base temporelle est le mois. Nous avons donc effectué des corrections pour ramener toutes nos séries dans cette base (voir annexe 2). Excepté ce traitement, nous n'avons pas eu à retravailler nos données, qui étaient de bonne qualité.

II.2. Statistiques descriptives

II.2.1. Analyse statistique et graphique

Différentes analyses ont été réalisées sur toutes nos variables, celles qui seront réellement utilisées pour la modélisation n'étant pas encore connues.

Un exemple de résultats sera présenté pour chacun des 3 types de série : une série de données sinistres (exemple : la fréquence mensuelle de sinistres en RC corporelle), une série de données portefeuille (exemple : le nombre de véhicules dans le foyer) et une série de données externes (exemple : les températures moyennes mensuelles). Les principales conclusions seront mentionnées ci-après.

Données sinistres :

L'étude visant à modéliser les fréquences mensuelles, nous sommes passés du nombre de sinistres à la fréquence de sinistres comme suit :

En RC Corporelle :

$$Fréquence\ \ mensuelle = \frac{Nombre\ de\ sinistres\ mensuels\ déclarés\ en\ RC\ corporelle}{Somme\ des\ années\ \ police\ en\ RC}$$

En RC matérielle :

$$Fréquence\ mensuelle = \frac{Nombre\ de\ sinistres\ mensuels\ déclarés\ en\ RC\ matérielle}{Somme\ des\ années\ \ police\ en\ RC}$$

La somme des années police RC est obtenue en sommant toutes les modalités de n'importe quelle variable, la garantie RC étant obligatoire.

En Dommages :

$$Fréquence\ mensuelle = \frac{Nombre\ de\ sinistres\ mensuels\ déclarés\ en\ Dommages}{Somme\ des\ années\ \ police\ en\ Dommages}$$

La somme des années police Dommages est obtenue en faisant :

AP[7] RC – nombre d'AP des assurés n'ayant pas souscrit la garantie Dommages.

En Incendie :

$$Fréquence\ mensuelle = \frac{Nombre\ de\ sinistres\ mensuels\ déclarés\ en\ Incendie}{Somme\ des\ années\ \ police\ en\ Incendie - Vol}$$

En Vol :

$$Fréquence\ mensuelle = \frac{Nombre\ de\ sinistres\ mensuels\ déclarés\ en\ Vol}{Somme\ des\ années\ \ police\ en\ Incendie - Vol}$$

La somme des années police Incendie –Vol est obtenue en faisant :

AP RC –AP des assurés n'ayant pas souscrit la garantie Incendie – Vol.

En Bris de glace :

$$Fréquence\ mensuelle = \frac{Nombre\ de\ sinistres\ mensuels\ déclarés\ en\ Bris\ de\ glace}{Somme\ des\ années\ \ police\ en\ Bris\ de\ glace}$$

avec : Somme des années police en Bris de glace = AP RC – AP des assurés n'ayant pas souscrit la garantie Bris de glace.

La première étape lors de l'analyse de séries temporelles est l'analyse exploratoire des données. Il s'agit de réaliser le tracé des séries, pour détecter visuellement la présence éventuelle de composantes : tendance, saisonnalité, cycle.

- → Le tracé peut être horizontal, et ne présenter que des fluctuations aléatoires autour d'un niveau constant : la série est stationnaire.
- → Le tracé peut évoluer à la hausse (ou à la baisse) au cours du temps : la série a une tendance.
- → Le tracé peut montrer une saisonnalité évidente etc.

D'où l'intérêt de l'analyse graphique de nos séries.

[7] AP : année police

Pour la série sur les «Fréquences de sinistres en RC Corporelle» :

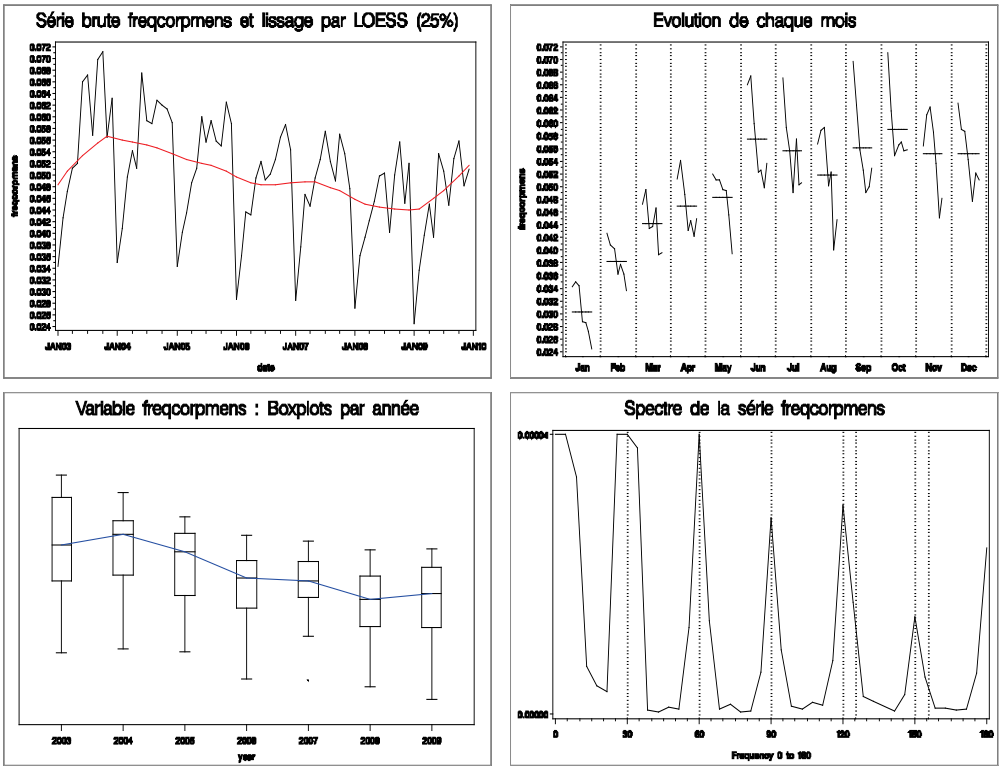


Figure 3 : Analyse graphique des fréquences en RC corporelle

On remarque la présence d'une forte saisonnalité de période 12, qui évolue régulièrement. Les creux observés sur les mois de janvier semblent s'être stabilisés sous la barre de 0,03. La série présente une tendance à la baisse depuis 2004, tendance qui repart à la hausse en 2009.

Nous avons effectué une analyse spectrale pour identifier les fluctuations de la série temporelle aux différentes fréquences.

Le spectre est un graphique représentant :

* en abscisse : les fréquences d'apparition des cyclicités (ces fréquences s'exprimant en radians pour la pulsation $\omega \in [0, \pi]$, ou en hertz pour la fréquence λ ou en unités de temps pour la période T)

* en ordonnée : les valeurs du spectre, qui peuvent être assimilées aux différentes variances des cyclicités de la fréquence ω, λ, T avec :

$\omega = \frac{2\pi}{T} = 2\pi\lambda$ et $\lambda = 1/T$, i.e. que la fréquence est l'inverse de la période.

Le périodogramme nous permet de mettre en évidence les composantes périodiques de la série temporelle. C'est le graphique du spectre des périodes.

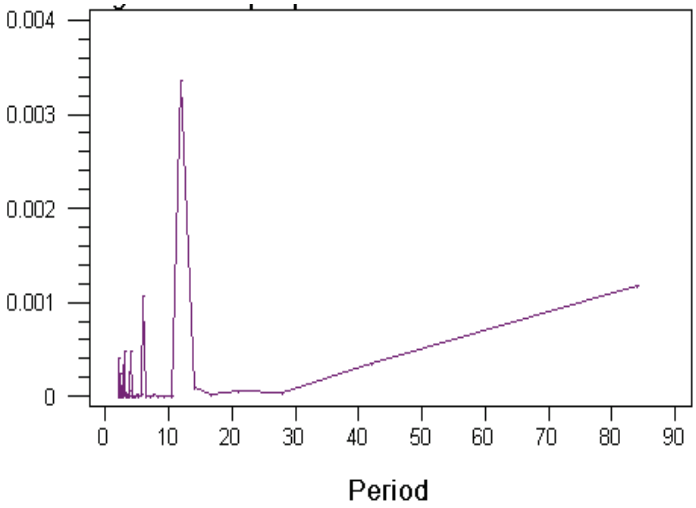


Figure 4 : Périodogramme des fréquences mensuelles en RC Corporelle

Le tracé du périodogramme nous confirme la présence de périodicités, avec des pics à l'origine. On observe le plus grand pic vers les basses fréquences, à 12, ce qui correspond à un cycle de 12 mois, puisque notre unité temporelle est le mois. La série n'est pas stationnaire.

Les résultats obtenus pour les fréquences mensuelles sont les suivants :

Pour la série sur les fréquences en RC matérielle, on observe une tendance linéaire à la baisse, avec une saisonnalité de période 12.
Pour la série sur les fréquences en Dommages, on observe des effets saisonniers, sans tendance marquée sur l'ensemble de la période.
Pour la série sur les fréquences en Vol partiel, on observe une tendance linéaire à la baisse avec une saisonnalité de période 12.
Pour la série sur les fréquences en Vol total, on n'observe qu'une tendance linéaire décroissante.
Pour la série sur les fréquences en Incendie, on n'observe ni tendance, ni saisonnalité, avec trois observations extrêmes entre 2003 et 2009.

Les graphiques de ces différentes séries sont disponibles en annexe 1.

Pour les coûts moyens, selon que les sinistres soient clos ou déclarés, nous avons obtenu les résultats suivants :

Evolution des coûts de sinistres clos avec comme année de référence 2003

Evolutions /2003	2004/2003	2005/2003	2006/2003	2007/2003	2008/2003	2009/2003
RC Corporelle	-6,0%	-9,7%	-9,7%	-16,2%	-7,7%	-10,1%
RC Matérielle	3,0%	4,1%	2,7%	3,1%	5,0%	9,0%
Dommages	1,7%	3,9%	3,4%	4,6%	4,2%	9,4%
Bris de glace	2,5%	8,0%	15,4%	21,3%	28,6%	32,1%
Incendie	19,5%	15,7%	36,2%	35,8%	46,9%	44,8%
Vol Total	21,9%	24,6%	26,7%	32,7%	35,3%	38,8%
Vol Partiel	5,7%	12,5%	20,9%	24,8%	30,2%	44,2%

En fin 2009, pour la garantie RC Corporelle, on note une baisse du coût moyen par rapport à 2003. Pour les autres garanties, les coûts moyens sont plutôt à la hausse.

Evolution des coûts de sinistres déclarés avec comme année de référence 2003

Evolutions /2003	2004/2003	2005/2003	2006/2003	2007/2003	2008/2003	2009/2003
RC Corporelle	26,6%	69,8%	-32,6%	-6,3%	89,5%	71,0%
RC Matérielle	8,5%	-4,0%	1,7%	4,1%	9,4%	20,9%
Dommages	2,8%	5,7%	5,2%	4,0%	3,6%	12,5%
Bris de glace	2,6%	8,2%	15,7%	21,6%	28,9%	32,2%
Incendie	4,4%	19,8%	41,6%	52,1%	60,8%	53,9%
Vol Total	23,6%	31,4%	28,9%	30,7%	43,3%	52,8%
Vol Partiel	5,8%	11,4%	19,3%	25,7%	27,1%	40,7%

Les évolutions des coûts de sinistres déclarés ne sont pas très représentatives de la réalité en RC Corporelle (hausse de 89% de la RC Corporelle en 2008 !!).

Evolution des coûts de sinistres clos d'une année sur l'autre

Evolutions N+1/N	2004/2003	2005/2004	2006/2005	2007/2006	2008/2007	2009/2008
RC Corporelle	-6,0%	-4,0%	0,1%	-7,2%	10,2%	-2,7%
RC Matérielle	3,0%	1,1%	-1,3%	0,3%	1,8%	3,8%
Dommages	1,7%	2,2%	-0,4%	1,1%	-0,4%	5,0%
Bris de glace	2,5%	5,3%	6,9%	5,1%	6,0%	2,7%
Incendie	19,5%	-3,2%	17,7%	-0,2%	8,2%	-1,5%
Vol Total	21,9%	2,2%	1,7%	4,8%	2,0%	2,6%
Vol Partiel	5,7%	6,4%	7,5%	3,2%	4,3%	10,7%

Les principales variations d'une année sur l'autre sont mentionnées ci-dessous :

Sur la garantie RC Corporelle, on observe une hausse de 10% en 2008.

Sur la garantie RC matérielle, on observe une hausse d'environ 4% en 2009.

Sur la garantie Dommages, on observe une hausse d'environ 5% en 2009.

Sur la garantie Bris de glace, on observe une hausse d'environ 7% en 2006.

Sur la garantie Incendie, on enregistre des fluctuations entre hausses et baisses.

Sur la garantie Vol total, on la plus forte variation a lieu entre 2003 et 2004 (+21,9 %)

Sur la garantie Vol partiel : hausse d'environ 11% en 2009

Evolution des coûts de sinistres déclarés d'une année sur l'autre

Evolutions N+1/N	2004/2003	2005/2004	2006/2005	2007/2006	2008/2007	2009/2008
RC Corporelle	26,6%	34,1%	-60,3%	39,0%	102,1%	-9,8%
RC Matérielle	8,5%	-11,6%	6,0%	2,3%	5,1%	10,5%
Dommages	2,8%	2,8%	-0,5%	-1,1%	-0,3%	8,6%
Bris de glace	2,6%	5,4%	7,0%	5,1%	6,0%	2,5%
Incendie	4,4%	14,7%	18,2%	7,5%	5,7%	-4,3%
Vol Total	23,6%	6,3%	-1,9%	1,5%	9,6%	6,6%
Vol Partiel	5,8%	5,2%	7,2%	5,4%	1,1%	10,7%

De cette analyse, il ressort que les coûts de sinistres déclarés fluctuent énormément pour la garantie RC corporelle. Des écarts sont également à signaler en termes d'évolution entre les coûts de sinistres déclarés et les coûts de sinistres clos pour les garanties RC matérielle, Dommages, Incendie et Vol total. Pour décider, pour chacune des garanties, quel serait idéalement le type de coût à modéliser, un coup d'œil sur les cadences de règlement est nécessaire.

Nous avons récupéré les triangles de liquidation ainsi que les cadences de règlements des garanties RC corporelle, RC matérielle et Matériels sur le périmètre Agents. Il faut savoir que le service Sinistres a regroupé sous l'appellation « Matériels » les garanties Dommages, Incendie-Vol et Bris de glace.

Les graphiques des cadences estimées sont représentés ci-dessous :

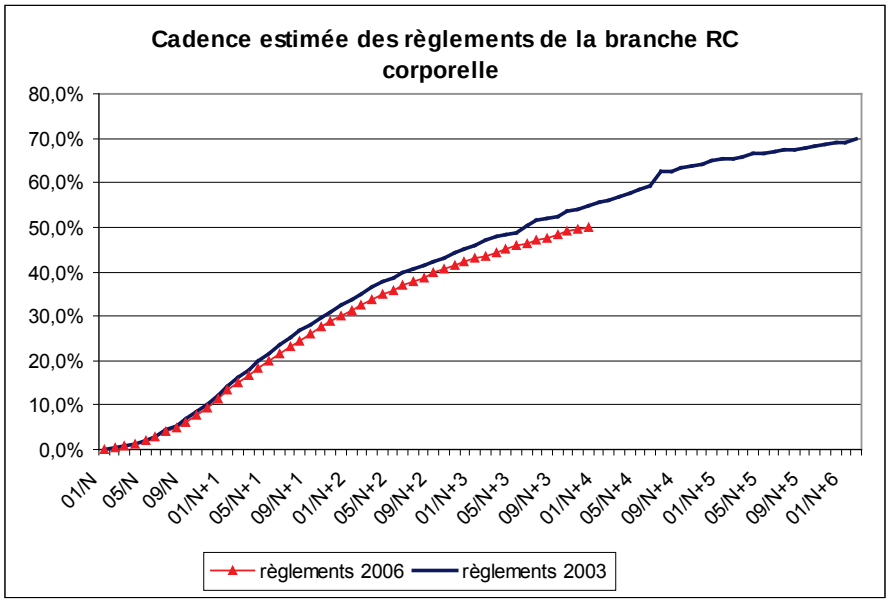


Figure 5 : Cadence estimée des règlements pour la branche RC corporelle

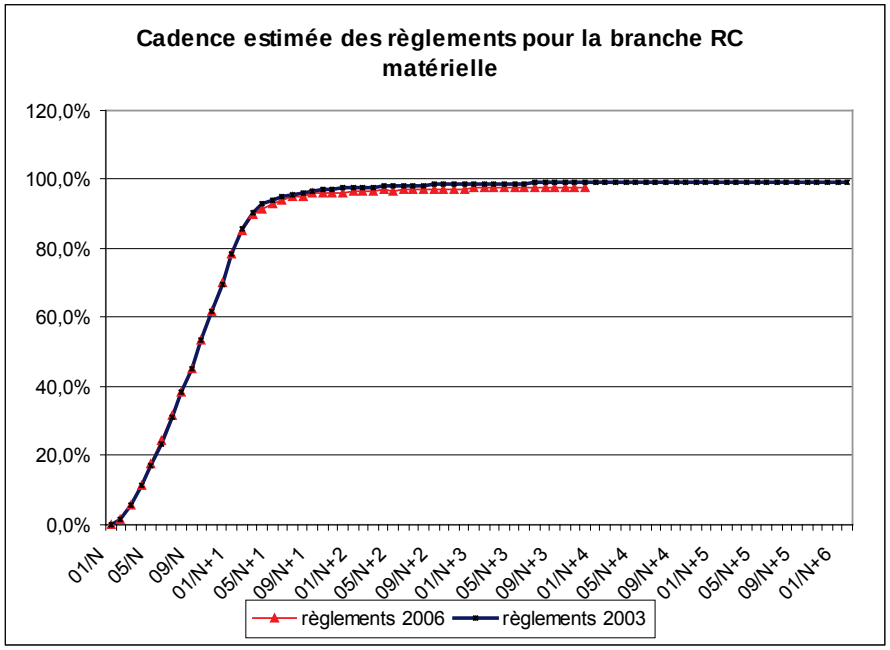


Figure 6 : Cadence estimée des règlements pour la branche RC matérielle

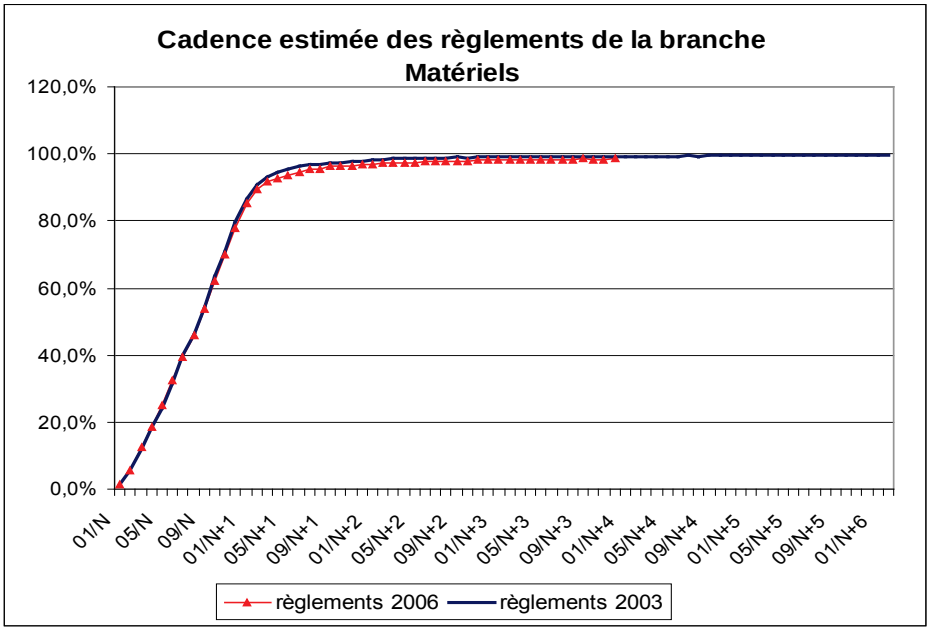


Figure 7 : Cadence estimée des règlements pour la branche Matériels

On constate que sur la branche RC corporelle, au bout de 6 ans, on n'a réglé que 70 % de la charge ultime estimée. Alors qu'en RC matérielle, au bout d'une année, on a déjà réglé 80 % de la charge ultime estimée. Pour les autres garanties, au bout d'une année, on a réglé plus de 86 % de la charge ultime estimée. Ces graphiques font apparaitre que la branche RC corporelle est à déroulement long, comme on pouvait s'y attendre : pour cette garantie, nous ne modéliserons que les coûts des sinistres clos. Pour les autres garanties (RC matérielle

comprise), on s'intéressera sans préférence particulière aux coûts des sinistres déclarés comme aux coûts des sinistres clos.

Données portefeuille :

Sur les données portefeuille présélectionnées, les analyses que l'on peut faire sont les suivantes :

La série sur le « Nombre de véhicules dans le foyer » :

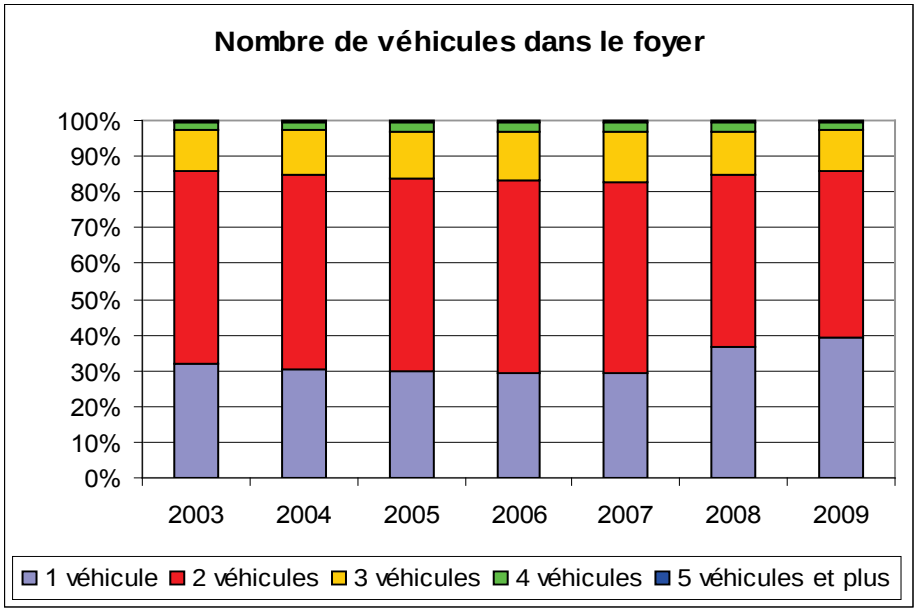


Figure 8 : Répartition annuelle des années police selon le nombre de véhicules

On observe une hausse de 7 % du nombre de foyers avec 1 véhicule, et une baisse de 7 % du nombre de foyers avec 2 véhicules entre 2003 et 2009. Néanmoins, les foyers avec 2 véhicules représentent toujours la population la plus importante (46 % du portefeuille en 2009).

La progression des foyers avec 1 véhicule était restée à la baisse (-1 % en 2004, -0,5 % en 2005, -0,7 % en 2006), avant de repartir à la hausse, notamment en 2008 avec + 7 %. Cette progression est contrebalancée par la baisse des foyers avec 2 véhicules.

Les séries sur les données portefeuille ne fluctuent pas énormément au cours du temps. Ce qui malgré un turn-over assez important chaque année peut s'expliquer par la part de marché importante d'AXA France.

Nous n'avons pas intégré tous les graphiques associés, du fait de leur grand nombre.

Données externes :

Sur les données externes, nous avons effectué les analyses suivantes :

Pour les variables météorologiques :

Exemple : La série sur les températures moyennes mensuelles

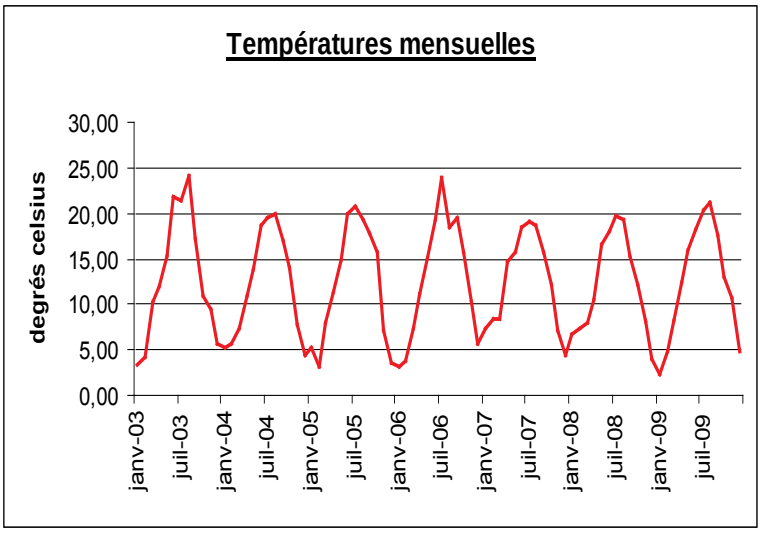


Figure 9 : Graphique des températures moyennes mensuelles

On observe un effet saisonnier annuel, avec une périodicité de 12. On remarque les pics de l'été (mois de juillet, août), avec deux pics plus marqués en août 2003 et août 2006. Ces deux derniers correspondent aux canicules qui ont touché la France. Vu qu'on est sur des moyennes mensuelles, les pics de températures sont moins marqués que les valeurs journalières qu'on a pu observés durant ces étés très chauds.

Le périodogramme obtenu nous confirme la tendance cyclique de période 12. La série n'est pas stationnaire.

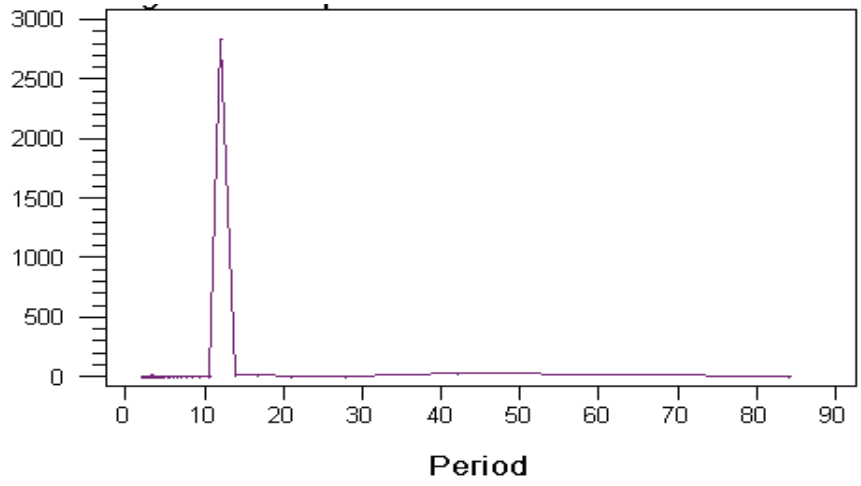


Figure 10 : Périodogramme de la série des températures moyennes

La série sur la pluviométrie : On observe une série plutôt fluctuante, avec une importante pluviométrie en mai 2007, et novembre 2009 (fortes intempéries dans le Pas de Calais), une baisse en février 2004 et avril 2007 dues à 2 sécheresses (dont la plus importante reste celle de 2004).

La série sur le gel : On observe un effet saisonnier de période 12, avec des pics en décembre 2007 et janvier 2009 (vagues de grand froid).

La série sur l'ensoleillement : on observe une saisonnalité annuelle. On retrouve des caractéristiques similaires à celle de la série sur les températures.

Pour les variables relatives aux véhicules et aux carburants :

Les séries sur les indices mensuels de prix :
Les indices des prix sont en base 100, avec comme année de référence 1998.
L'IPC sur les carburants et lubrifiants peut être ajusté à une courbe polynomiale avec un assez bon R^2 de 0,66
L'IPC des pièces et accessoires de véhicules suit une tendance linéaire avec un R^2 égal à 0,98.
L'IPC des réparations de véhicules personnels suit également une tendance linéaire avec un R^2 égal à 0,99.

Les séries sur les volumes mensuels de livraison de carburants :
On observe une saisonnalité annuelle, avec tendance à la baisse pour la livraison de Super (pics en juillet-août : départs en vacances). Pour la livraison de Gazole, la série est fluctuante, sans saisonnalité, avec une tendance à la hausse qui se stabilise en 2009.
On retrouve la saisonnalité et l'absence de tendance dans la série des livraisons confondues de Gazole et de Super.

Les séries sur la consommation annuelle de carburants :
On constate une hausse de 10 % sur la période 2003 – 2007, suivie d'une légère baisse (-0,3 %) en 2008, pour la consommation de Gazole.
La consommation d'essence et de supercarburants enregistre une baisse de 26% sur la période 2003-2008.

Les séries sur les moyennes mensuelles des prix des carburants :

Le Super plombé n'est plus commercialisé depuis fin 2006.

On observe une tendance très forte des prix à la hausse, jusqu'en juillet 2008 où on enregistre une forte baisse des prix des carburants due à la crise économique (chute de la demande de pétrole). Puis on enregistre une reprise à la hausse début 2009 avant une stabilisation en novembre-décembre 2009.

Les séries sur le parc annuel et le nombre mensuel d'immatriculations de véhicules :

Le parc annuel connait une augmentation annuelle faible (+1,15 % en 2004, + 0,67 % en 2005, + 1 % en 2006, + 0,9 % en 2007, + 0,49 % en 2008). Tandis que les immatriculations de véhicules (neufs et d'occasion) sont très volatiles.

Pour les variables liées aux comportements des assurés :

Les séries sur les parcours mensuels sur les différents axes routiers (en Milliards de véhicules par km) :

Excepté sur les autoroutes et voies rapides urbaines, on observe une saisonnalité annuelle avec des pics en juillet. La tendance à la hausse jusqu'en 2007 est suivi d'une baisse en juillet 2008 (due à la flambée des prix des carburants et à la baisse du pouvoir d'achat). On constate une reprise à la hausse en 2009 (encouragée par la baisse des prix des carburants).

Sur les autoroutes et voies rapides urbaines, on a série plutôt stable jusqu'à une forte hausse en janvier 2006. Début 2008, une baisse s'est effectuée même si on reste au-dessus des valeurs d'avant 2006.

Les séries sur les vitesses moyennes de jour et de nuit sur différents axes routiers :

On constate une baisse générale de toutes les vitesses sur les axes routiers, avec une légère remontée fin 2009 de la vitesse de jour sur les RN à 2 ou 3 voies, et de la vitesse de nuit sur les RN traversées d'agglomération, sur les entrées et sorties d'agglomération et sur les RN à 2 ou 3 voies.

Les séries sur les nombres d'accidents corporels, de victimes blessées et tuées :

Sur la période 2003 – 2009, on constate une baisse de 24 % du nombre d'accidents corporels, de 25,6 % du nombre de victimes décédées et de 27,6 % du nombre de victimes blessées.

Il ressort de ces analyses que nos séries ne sont pas stationnaires : on observe des effets saisonniers annuels, des tendances à la hausse ou à la baisse sur la période d'étude. Nous ne ferons de transformations pour les rendre stationnaires que si le type de modèle utilisé nous l'impose.

I.2.2. Etude des corrélations

La modélisation que nous souhaitons effectuer porte sur les fréquences et les coûts moyens des sinistres. Donc les corrélations qui nous intéressent sont celles entre les données sinistres (Fréquences / coûts par UP) et les autres données (externes / portefeuille). Les données sinistres étant des variables quantitatives, les indicateurs que nous pouvons utiliser sont les coefficients de corrélation de Pearson et de Spearman.

Le coefficient de corrélation de Pearson :

Il permet de mesurer l'intensité de la liaison entre deux variables continues, pas trop éloignées d'une distribution gaussienne. La liaison recherchée ici est une relation affine. On calcule

ρ = **Erreur ! Signet non défini.** $\frac{\mathrm{cov}(r_x, r_y)}{\sigma_{rx}\sigma_{ry}}$

Ce coefficient est compris en -1 et +1. Plus il est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte.

Il n'est pas sensible aux unités des variables mais est extrêmement sensible à la présence de valeurs aberrantes ou extrêmes.

Le coefficient de corrélation de Spearman :

Le test du coefficient de corrélation de Spearman correspond à l'équivalent non paramétrique du test basé sur le coefficient de corrélation de Pearson. C'est un test sur les rangs, qui ne dépend pas de la normalité de la série, ni de la taille de l'échantillon testé. La forme de la liaison, linéaire ou non, ne peut pas être observée. En revanche, l'existence d'une liaison monotone même non linéaire est très bien perçue. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte. Il est insensible aux valeurs aberrantes, et peut être appliqué dans le cas de variables continues, discrètes ou ordinales.

C'est ce coefficient que nous utiliserons pour comparer les corrélations avec le test suivant :

H_0 : le coefficient de corrélation est nul, il n'y a pas de corrélation des rangs

H_1 : le coefficient est significativement différent de 0.

Nous citons ci-dessous, pour chaque garantie, quelques-unes des corrélations les plus significatives avec les variables explicatives dont nous disposons : une p-value < 1 % et un coefficient de Spearman > |0,4|.

Pour les séries de fréquences mensuelles :

Variables explicatives potentielles	RC corporelle	RC matérielle	Dommages	Bris de glace	Incendie	Vol partiel	Vol total
Températures	0,45			0,66	0,46		
Ensoleillement				0,61	0,41		
Gel	-0,43			-0,63			
Vitesse de jour sur les RD	0,50	0,43				0,78	0,85
Vitesse de jour sur les autoroutes de dégagement						0,49	
Vitesse de nuit sur les RN à 2 ou 3 voies					-0,49	0,40	0,51
Vitesse de nuit sur les artères en agglomération						0,53	0,66
Volume de Super et Gazole	0,63	0,64	0,49	0,67	0,45	0,48	
Volume de Super	0,60	0,44				0,82	0,73
Volume de Gazole		0,41	0,52	0,48	0,52		
Parcours mensuel - Réseau non concédé total 20				0,57	0,54		
Parcours mensuel autoroutes interurbaines 21				0,52	0,53		-0,45
Parcours mensuel - Autoroutes et voies rapides urbaines 22		0,40	0,41	0,46	0,50		
Parcours mensuel - Routes nationales interurbaine à caractéristiques autoroutières 23				0,45	0,51		-0,54

Parcours mensuel - utres routes nationales 24	0,54			0,67	0,45		
Parcours mensuel moyen - Réseau national 25				0,54	0,51		
Parcours mensuel moyen autoroutes non concédées 26				0,54	0,58		-0,42
Parcours moyen mensuel - total des autoroutes 27				0,53	0,51		-0,40
Parcours mensuel moyen - routes nationales 28				0,57	0,51		
Ancienneté de carte grise de 3 ans	0,45				-0,40	0,76	0,87
Age 36-45 ans	-0,44					-0,77	-0,90

Pour les séries de coûts de sinistres clos :

Variables explicatives potentielles	RC corporelle	RC matérielle	Dommages	Bris de glace	Incendie	Vol partiel	Vol total
Vitesse de jour sur les autoroutes de dégagement	0,43		-0,64	-0,74	-0,52	-0,75	-0,67
Vitesse de jour sur les RD			-0,69	-0,89	-0,72	-0,91	-0,82
Vitesse de nuit sur les RN 2-3 voies	0,45			-0,55		-0,47	-0,50
Vitesse de nuit sur les artères en agglomération		-0,41	-0,52	-0,71	-0,61	-0,59	-0,70
Volume de Super et Gazole							
Volume de Super			-0,61	-0,79	-0,57	-0,82	-0,68
Volume de Gazole				0,53	0,49		0,51
IPC réparation des véhicules personnels		0,42	0,60	0,99	0,83	0,87	0,94
IPC pièces et accessoires		0,42	0,60	0,99	0,83	0,87	0,94

Nombre de victimes décédées			-0,64	-0,94	-0,82	-0,90	-0,88
Accidents corporels			-0,65	-0,95	-0,79	-0,91	-0,85
Parcours mensuel - Réseau non concédé total 20		0,43			0,45		0,44
Parcours mensuel – Autoroute interurbaine 21		0,46		0,46	0,52		0,52
Parcours mensuel - Autoroutes et voies rapides urbaines 22							
Parcours mensuel - Route nationale interurbaine à caractéristiques autoroutières 23		0,44		0,51	0,55		0,57
Parcours mensuel - autres routes nationales 24							
Parcours mensuel moyen - Réseau national 25		0,41			0,46		0,43
Parcours mensuel moyen - autoroutes non concédées 26	-0,40	0,43		0,48	0,51		0,53
Parcours moyen mensuel - total des autoroutes 27					0,46		0,43
Parcours mensuel moyen - routes nationales 28		0,43			0,42		0,40
Age 18-25			-0,69	-0,98	-0,79	-0,93	-0,90
Age 26-35			-0,69	-0,98	-0,79	-0,93	-0,90
Age 36-45			0,69	0,98	0,79	0,93	0,90
Usage tournées 5			-0,69	-0,98	-0,79	-0,93	-0,90
Usage 4			0,69	0,95	0,80	0,89	0,88
Usage 3			0,62	0,95	0,80	0,89	0,88

Pour les séries de coûts de sinistres déclarés :

Variables explicatives potentielles	Dommages	Bris de glace	Incendie	Vol partiel	Vol total
Vitesse de jour sur les RD	-0,51	-0,88	-0,88	-0,73	-0,60
Volume de Super et Gazole					
Volume de Super	-0,56	-0,79	-0,81	-0,60	-0,46
Volume de Gazole		0,53		0,58	0,53
IPC réparation des véhicules personnels	0,47	0,99	0,84	0,90	0,80
IPC pièces et accessoires	0,47	0,99	0,84	0,90	0,80
Parcours mensuel - Réseau non concédé total 20				0,54	0,55
Parcours mensuel - Réseau non concédé total 21		0,47		0,61	0,62
Parcours mensuel - Autoroutes et voies rapides urbaines 22				0,48	
Parcours mensuel - Route nationale interurbaine à caractéristiques autoroutières 23		0,53		0,64	0,64
Parcours mensuel - autres routes nationales 24					
Parcours mensuel moyen - Réseau national 25				0,54	0,54
Parcours mensuel moyen - autoroutes non concédées 26		0,49		0,63	0,60
Parcours moyen mensuel - total des autoroutes 27				0,54	0,53

Il est fondamental de noter qu'une corrélation significative ne signifie aucunement qu'il existe une relation de cause à effet entre les deux variables. En réalité, les deux variables peuvent être corrélées à une troisième variable non mesurée, et dont dépendent les deux autres. La corrélation n'implique pas la causalité, mais la causalité implique la corrélation. L'absence de corrélation est donc une preuve de l'absence de causalité. Nous n'entrerons pas dans le détail de la causalité, sujet assez délicat, et nous nous contenterons de ce postulat pour poursuivre cette étude.

La démarche pour la suite est la suivante : nous commençons par réaliser des modèles sans variables explicatives ; ensuite nous intégrons dans ces modèles les variables les plus corrélées avec la série cible. Après un test sur la stabilité des modèles obtenus, nous comparons ceux qui réussissent la validation selon des critères (RMSE, MAPE, MPE) et nous effectuons des prévisions sur le premier semestre 2010 pour décider du meilleur modèle à conserver.

III. MODELISATION

Nous voulons étudier des phénomènes qui évoluent dans le temps, c'est-à-dire les décrire, les expliquer, les contrôler ou les prédire. Il nous faut trouver un modèle déduit de nos observations et qui ait du sens. Avant toute analyse, nous allons commencer par décrire certains éléments mathématiques que nous utiliserons dans la suite.

III.1. Quelques rappels mathématiques

Soit $(\Omega, \mathcal{A}, \mathbb{P})$, un espace probabilisé avec Ω l'espace des évènements, $\mathcal{A}$ une tribu adaptée à Ω et IP une mesure de probabilité définie sur $\mathcal{A}$. Soient $(T, \mathcal{T})$ et $(\Omega', \mathcal{A}')$ deux espaces mesurables ; un processus stochastique est une application Y définie sur $\Omega \times T$, à valeurs dans Ω', associant au couple (ω, t) la réalisation $Y(\omega,t)$, encore notée $Y_t(\omega)$ telle que pour t fixé appartenant à T, Y est une variable aléatoire (v.a.) sur $(\Omega, \mathcal{A})$.

Par extension, on écrira un processus sous la forme d'une suite de v.a. indicées par t, notée $(Y_t, t \in T)$ ou plus simplement (Y_t). La loi du processus est l'image IP^Y de IP par Y. Lorsque Ω' est IR (respectivement IR^m), le processus est dit réel unidimensionnel ou univarié (respectivement multidimensionnel de dimension m ou m-varié) ; si Ω' est fini ou dénombrable, on parle de processus à valeurs discrètes.

Une série temporelle $(Y_t)_{t=1,\ldots,T}$ est la réalisation particulière d'un processus stochastique $(Y_t, t \in T)$. Dans cette étude, les séries que nous étudierons seront toutes réelles et univariées. Nous ne considèrerons que des processus discrets, c'est-à-dire où T est de la forme {1, …, n}.

Les différents types de modélisation que nous allons appliquer à nos séries temporelles sont celles expliquées ci-dessous.

III.2. Modélisation ARMA

Les modèles ARMA ne sont applicables qu'à des processus stationnaires. Il faut que toute portion de la trajectoire de nos observations fournisse des informations sur la loi de la variable aléatoire, et que des portions différentes, mais de même longueur, fournissent les mêmes

indications. Il nous faut des processus stables dans le temps. C'est ce qui nous amène à définir la notion de stationnarité.

Définition de la stationnarité : (nous n'aborderons que le cas unidimensionnel)

Un processus aléatoire $Y = (Y_t, t \in T)$ est dit strictement stationnaire ou fortement stationnaire si : $P(Y_{t1},\ldots, Y_{tk}) = P(Y_{t1+h}, \ldots, Y_{tk+h})$, $\forall\, k \geq 1$, $\forall\, (t_1, \ldots, t_k) \in T$, $\forall\, h$ tel que $(t_1+h, \ldots, t_k+h) \in T$.

Y est faiblement stationnaire ou stationnaire du second ordre si sa moyenne m(t) et sa fonction d'autocovariance cov (Y_s, Y_t) sont invariantes par translation dans le temps :

$\forall\, t \in T$, $\quad E(Y_t) = m(t) = m$, $V(Y_t) = \sigma^2$.

$\forall\, t, h \in T$, $\quad cov\,(Y_t, Y_{t+h}) = \gamma(h)$.

L'un des processus les plus courants est le bruit blanc (noté b.b.). Un bruit blanc fort est une suite $(\varepsilon_n)_{n \in Z}$ de v.a. réelles indépendantes, de même loi, centrées, et telles que :

$0 < \sigma^2 = E\,(\varepsilon_n^2) < \infty$, $n \in Z$.

Si les variables ne sont plus indépendantes, mais juste orthogonales, c'est-à-dire que $cov\,(\varepsilon_n, \varepsilon_m) = 0$, $n \neq m$, alors on obtient un bruit blanc faible.

Un bruit blanc fort est strictement stationnaire, alors qu'un bruit blanc faible est faiblement stationnaire.

La stationnarité du second ordre est bien plus facile à étudier et à vérifier que la stationnarité stricte. Son importance pratique tient surtout aux problèmes de prédiction ou de régression. En effet, on se limite souvent à des critères de moindres carrés pour avoir des estimateurs calculables. Cela signifie alors utiliser des prédicteurs linéaires optimaux dont le calcul ne fait pas intervenir dans sa totalité la structure probabiliste du processus Y observé, mais seulement la géométrie (angles et longueurs) de la suite (Y_k) considérée comme suite de vecteurs dans l'espace de Hilbert $L^2(\Omega\,; P)$. Or, cette géométrie ne dépend que des moments d'ordre 2 de Y; la notion naturelle de stationnarité est donc l'invariance de ces moments d'ordre 2 par translation dans le temps.

Dans le cas gaussien, les propriétés du second ordre déterminent complètement les distributions jointes. Les processus gaussiens stationnaires sont définis de manière unique par

leur fonction d'autocovariance. Dans le cas d'un processus non gaussien, ce n'est plus le cas mais cette fonction reste la façon la plus simple de modéliser la dépendance du processus, en ne s'intéressant qu'aux moments d'ordre 2.

La fonction d'autocovariance notée γ est définie par :

$\forall$ s, t$\in$T, $\gamma(s, s+t) = cov(Y_s, Y_{s+t}) = \gamma(t)$ (car ne dépend que de t).

Pour Y un processus réel, γ est une fonction paire.

La fonction d'autocorrélation $\rho(.)$ est une normalisation de la fonction d'autocovariance et satisfait la condition supplémentaire $\rho(0) = 1$.

L'estimateur de la fonction d'autocorrélation et l'estimateur de la fonction d'autocorrélation partielle permettront de faire une première sélection parmi les nombreux modèles stationnaires susceptibles de représenter la dépendance des données. Par exemple, une ACF (Autocorrelation Function) empirique très proche de 0 suggère qu'un modèle adapté aux données pourrait être un bruit blanc fort.

On considère le modèle sous sa forme additive : $Y_t = f(t) + s(t) + \varepsilon_t$; t$\in$T

où f = tendance : fonction déterministe

s = saisonnalité : fonction déterministe et périodique

ε_t = résidu : processus stochastique centré stationnaire.

On veut estimer et éliminer f et s, de façon à ne garder que l'observation de la partie stationnaire (ε_t).

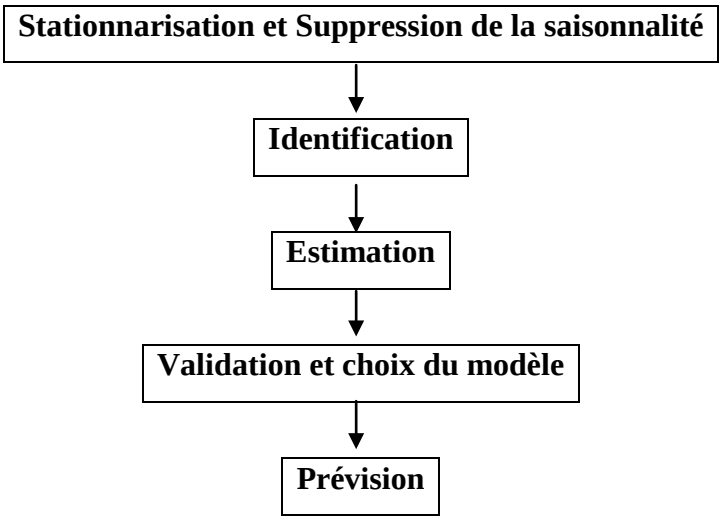
Dans tout ce qui suit, on supposera le processus Y centré. Il suffira dans le cas où Y est non centré de moyenne μ ($E(Y_t) = \mu$; $\forall$ t), de poser $Z_t = Y_t - \mu$ et d'appliquer nos résultats à $Z = (Z_t; t \in T)$.

On dispose de deux types de méthodes :

Soit estimer la tendance et la saisonnalité et les soustraire de la série considérée ;

Soit « différencier » la série (c'est la méthode de Box-Jenkins).

Nous avons opté pour la deuxième approche, celle de la différenciation. Les étapes que nous allons ensuite suivre sont les suivantes :



Stationnarisation et Suppression de la saisonnalité :

On définit l'opérateur retard (noté L pour Lag ou B pour Backward) comme suit : Soit un processus stochastique $(Y_t,\ t \in Z)$. Alors $BY_t = Y_{t-1},\ \forall\ t \in Z$. Cet opérateur définit une application qui, à toute variable Y_t associe la variable retardée Y_{t-1}.

On introduit l'opérateur $\nabla = 1 - B$ et de façon générale $\nabla^j = \nabla\ (\nabla^{j-1})$, avec $\nabla^0 = I$.

Si on applique ∇^k à une fonction à tendance polynômiale d'ordre k, on la réduit à une constante. En pratique, k est petit.

Pour les séries avec une saisonnalité de période s, nous utilisons l'opérateur

$\nabla_S = 1 - B^s$: $\nabla_S Y_t = Y_t - Y_{t-s}$. En appliquant ∇_S à notre modèle $(*)$, on obtient :

$$\nabla_S (f_t + s_t + \varepsilon_t) = f_t - f_{t-s} + \varepsilon_t - \varepsilon_{t-s}.$$

On se ramène ainsi à une estimation de la tendance.

On a donc effectué ces transformations pour obtenir une série stationnaire, en particulier sans tendance ni saisonnalité.

Identification :

L'identification conduit à déterminer les ordres p et q des parties autorégressives et moyenne mobile du modèle et les ordres de différenciation éventuelles.

Un processus stationnaire peut être modélisé par la classe des ARMA (p,q).

Un processus $(Y_t, t \in Z)$ est un processus ARMA (p,q) si le processus est stationnaire et si $\forall\, t \in Z$, $\Phi(B)Y_t = \Theta(B)\varepsilon_t$ où ε_t est un b.b. de variance σ^2, Φ et Θ sont des polynômes de degrés p et q respectivement définis par :

$\Phi(z) = 1$ - **Erreur ! Signet non défini.** $\varphi_1 z$ - … - $\varphi_p z^p$ $\quad (\varphi_{p \neq} 0)$

$\Theta(z) = 1$ - **Erreur ! Signet non défini.** $\theta_1 z$ - … - $\theta_p z^p$ $\quad (\theta_{p \neq} 0)$

avec les θ_i, φ_i qui sont des réels tels que Φ et Θ n'ont pas de racines communes.

$(Y_t, t \in Z)$, processus stationnaire, est un processus autorégressif d'ordre p noté AR(p) si $\Theta(z) = 1$.

$(Y_t, t \in Z)$, processus stationnaire, est un processus moyenne mobile d'ordre q noté MA(q) si $\Phi(z) = 1$.

Estimation :

L'estimation est la phase la plus mécanique de la démarche. Les logiciels de statistique utilisent en général soit la méthode du maximum de vraisemblance (ML), soit la méthode des moindres carrés conditionnels (CLS), soit la méthode des moindres carrés non conditionnels (ULS).

Avant de passer à la modélisation, si on a veut s'assurer qu'on a suffisamment différencié notre série, on pourra effectuer des tests de racine unité. La méthode que nous utilisons est celle Dickey et Fuller (voir annexe 5).

Choix du modèle :

Pour le choix définitif parmi les modèles ayant passé le cap de la vérification, il existe deux types de critères, qui ne sont pas très bien fondés statistiquement (au sens mathématique) mais qui donnent de bons résultats empiriquement. Il s'agit des critères d'information. La méthode consiste à choisir le modèle en se basant sur une mesure de l'écart entre la vraie loi (inconnue) et celle du modèle proposé. La mesure habituellement utilisée est l'information de Kullback. On choisit le modèle donnant la plus petite valeur de l'estimation de l'information de Kullback. Parmi les critères d'estimation basés sur l'information de Kullback, on en citera deux qui sont utilisés avec le logiciel SAS, à savoir :

1) le critère d'information d'Akaike (Akaike, 1974; Harvey, 1981), basé sur la statistique

AIC (p,q) = – 2 ln ($L_X(\hat{p},\hat{q},\hat{\sigma}^2)$) + 2(p + q + 1)

où L_X est la vraisemblance de Y lorsque le processus est gaussien.

2) le critère bayésien de Schwarz (Akaike, 1978), utilisé dans le cas d'un processus ARMA(p,q) centré causal et inversible. Il est noté SBC (Schwarz Bayesian Criterion) ou BIC (Bayesian modification of the AIC criterion). Il a été introduit par Akaike pour corriger la tendance de l'AIC à surestimer le nombre de paramètres. Il est basé sur la statistique :

SBC(p,q) = BIC(p,q) = – 2 ln ($L_X(\hat{p},\hat{q},\hat{\sigma}^2)$) + 2(p + q + 1) ln *n*;

où n est le nombre d'observations.

On choisira le couple (p,q) qui minimise ces statistiques.

Pour des modèles ARIMA saisonniers, on obtient des modèles multiplicatifs du type :

$(1 - B)^d (1 - B^s)^D \Phi(B) \phi(B^s) Y_t = \Theta(B)\theta(B^s)\varepsilon_t$

où $\phi(B)$ est un polynôme de degré P, $\Phi(B)$ un polynôme de degré p, $\theta(B)$ un polynôme de degré Q et $\Theta(B)$ un polynôme de degré q, et pour $|z| < 1$.

Un processus (Y_t) satisfaisant cette équation est appelé processus $SARIMA[(p,d,q)(P,D,Q)]_s$ ou $SARIMA_s[(p,d,q)(P,D,Q)]$.

Validation:

La validation porte sur

- l'analyse de la qualité statistique du modèle estimé (tests sur les estimateurs, qualité globale de l'ajustement)
- l'analyse des résidus d'estimation : si l'identification du modèle est correcte, les résidus sont des réalisations de bruits blancs.

Tests sur les résidus:

Dans cette étape, on modélise la série du bruit estimé, ou résidus.

- S'il n'y a pas de dépendance entre ces résidus, ils peuvent alors être considérés comme les observations de variables aléatoires indépendantes. Il ne reste donc qu'à en estimer la moyenne et la variance.
- S'il y a une dépendance significative entre ces résidus, alors on doit chercher à modéliser les bruits par une série stationnaire plus complexe (grâce à la théorie des processus

stationnaires), qui tienne compte de ce facteur. Cette dépendance permet en particulier d'utiliser les observations passées pour prédire des valeurs futures.

On propose des tests (paramétriques et non-paramétriques) pour vérifier si les résidus sont des valeurs observées de variables aléatoires indépendantes et identiquement distribuées. Ces tests nous permettent également de vérifier la validité du modèle. Ils sont détaillés en annexe 4.

Prévision :

Une fois le modèle validé, le nouvel enjeu est de pouvoir prédire les valeurs futures $(Y_t, t \geq n+1)$ de Y à partir des valeurs observées $Y_1, \ldots, Y_n$.

Etant donné un sous-espace fermé M de $L^2(\Omega)$, la meilleure prédiction dans M de $Y_{n+h}(\in L^2)$ est définie comme l'élément de M ayant la plus petite distance au sens des moindres carrés de Y_{n+h}. On la note $\hat{Y}_{n+h}$ et on l'appelle prédiction au sens des moindres carrés de Y_{n+h} dans M. Ainsi :

$\| Y_{n+h} - \hat{Y}_{n+h} \|^2 = \inf_{z \in M} \| Y_{n+h} - Z\|^2 = \inf_{z \in M} E[(Y_{n+h} - Z)^2]$ et $\hat{Y}_{n+h} = P_M(Y_{n+h})$: projection de Y_{n+h} sur M.

Bien sûr, ce n'est pas la seule définition possible de « meilleure prédiction », mais pour les processus du second ordre, cette définition permet d'introduire une théorie de la prédiction qui est simple et utile en pratique.

On s'intéresse au prédicteur dit «linéaire » $P_{[1, Y_1, \ldots, Yn]}(Y_{n+h})$ de Y_{n+h}, étant donné que le calcul de $P_{M(Y_1, \ldots, Y_n)}(X_{n+h})$ s'avère en général difficile à effectuer et est donc peu utilisé en pratique.

Ainsi, sous certaines conditions d'inversibilité, le meilleur prédicteur linéaire $\hat{Y}_{n+h}$ de Y_{n+1} en termes de $Y_1, \ldots, Y_n$ est : $\hat{Y}_{n+h} = \sum_{i=1}^{n} \phi_{ni} Y_{n+1-i}$, n= 1, 2, …

où $\phi_n = {}^t(\phi_{n1}, \ldots, \phi_{nn}) = {}^t(\gamma(1), \ldots, \gamma(n))$

L'erreur de prédiction est égale à $E[(Y_{n+h} - \hat{Y}_{n+h})^2]$.

Cette erreur de prédiction tend vers γ(0), la variance de Y, d'autant plus vite que γ (h) décroît rapidement en fonction de h. Cela signifie que l'information apportée par le présent pour prédire un futur "lointain" est à peu près nulle.

Pour n fixé, les erreurs de prédiction Y_{n+h} - $\hat{Y}_{n+h}$, h=1, …, 2 ne sont pas non-corrélées. En fait, d'après $Y_{n+h} - \hat{Y}_{n+h} = \sum_{j=0}^{h-1} \psi_j \, \varepsilon_{n+h-j}$, la covariance entre les erreurs au pas *k* et *h* vaut :

$$E\left[(Y_{n+h} - \hat{Y}_{n+h})(Y_{n+k} - \hat{Y}_{n+k})\right] = \sigma^2 \sum_{j=0}^{h-1} \psi_j \psi_{j+k-h}, \; k \geq h.$$

On peut obtenir des intervalles de confiance pour prédiction à un pas donné dans le cas gaussien. Ainsi à 95% :

$P(Y_{n+h} \in [\hat{Y}_{n+h} - 1,96\ \sigma(h)\ ;\ \hat{Y}_{n+h} + 1,96\ \sigma(h)] = 0,95$

où σ^2 est l'erreur quadratique moyenne au pas *h*.

III.3. Modèle de régression avec erreurs ARMA

Ce type de modèle est une des généralisations du modèle ARMA.

Dans le modèle de régression classique, $Y_t = \sum_{j=1}^{n} b_j X_{nj} + \varepsilon_t$ (où les X_{nj} sont non aléatoires), on fait souvent l'hypothèse que ε_t est un bruit blanc. Une autre hypothèse que l'on peut faire est que ε_t est un processus autorégressif (AR) ou, de façon plus générale, que ε_t est un ARMA(p,q).

L'équation devient :

$$Y_t = \sum_{j=1}^{n} b_j X_{nj} + \frac{\Theta(B)}{\Phi(B)}\eta_t, \quad t = 1,\ldots, T \text{ avec } \eta_t \text{ un b.b. de variance } \sigma^2.$$

Dans les variables explicatives, nous pouvons introduire des « variables d'intervention ». Les séries chronologiques sont souvent perturbées par des évènements spéciaux. Lors de la modélisation d'une série, l'analyse d'intervention (Box et Tiao (1975)) prend en compte ces interventions extérieures et fournit une mesure de l'impact de celles-ci sur la série.

La variable d'intervention représente l'effet d'une intervention extérieure à la date t', mis sous la forme d'une variable déterministe qui prend pour valeur 1 ou 0 suivant la présence ou l'absence de l'intervention. Cette variable est en général modélisée de deux manières :

→ soit par une variable en forme de saut ou d'escalier :

$$S_t^{(T)} = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases} \quad \textbf{Erreur ! Signet non défini.}$$

(La variable d'intervention est alors destinée à rendre compte de l'influence d'un phénomène commençant à la date T, par exemple un changement de réglementation.)

→ soit par une variable en forme d'impulsion :

$$P_t^{(T)} = \begin{cases} 0, & t \neq T \\ 1, & t = T \end{cases} \quad \textbf{Erreur ! Signet non défini.}$$

(La variable d'intervention est alors destinée à rendre compte de l'influence sur Y_t d'un phénomène ayant lieu à la date T uniquement, par exemple une grève.)

On notera que : $S_t^{(T)} - S_{t-1}^{(T)} = P_t^{(T)}$. Il n'y a pas de nombre limité d'interventions : la série chronologique peut être perturbée par k interventions.

On peut également prendre en compte la saisonnalité dans nos modèles grâce à des variables d'intervention. Des variables, codées en binaire (X_1 vaut 1 tous les mois de janvier et 0 sinon, …, X_{11} vaut 1 tous les mois de novembre et 0 sinon), ont ainsi été rajoutées dans nos modèles.

Pour cette modélisation, les procédures AUTOREG et ARIMA de SAS nous seront utiles.

La procédure AUTOREG nous permet de faire des régressions linéaires simples d'une variable dépendante Y_t sur des variables explicatives X_{1t}, X_{2t}, … pour des séries temporelles. Dans les régressions sur de telles séries, les erreurs sont très rarement indépendantes et identiquement distribuées. La procédure AUTOREG offre un ensemble de tests sur les erreurs et des méthodes d'estimation dans ce cas. On peut modéliser des erreurs AR de tout ordre, voire même utiliser une procédure de sélection pour sélectionner l'ordre "automatiquement"; le diagnostic de l'autocorrélation repose essentiellement sur le test de Durbin-Watson généralisé (voir annexe 4).

Le modèle proposé pour la série est de la forme : $Y_t = \sum_{j=1}^{n} b_j X_{nj} + \varepsilon_t$, $1 \leq n \leq N$, avec ε_t qui est un processus AR(p), $p \in$ IR*.

De même, la procédure ARIMA permet d'introduire des variables exogènes. Et le modèle obtenu peut être vu comme un modèle de régression multiple où on a spécifié que les perturbations suivaient un modèle AR, ou MA, ou ARMA.

III.4. Modélisation UCM

Une autre approche que nous avons abordée au cours de cette étude a été celle des modèles à composantes inobservables ou UCM (Unobserved Components Models).
Les modèles UCM allient la flexibilité des modèles ARIMA à la simplicité d'usage et d'interprétation des « modèles de lissage ». En effet, chacune des composantes du modèle peut être directement interprétée, tandis qu'avec les modèles SARIMA, des difficultés d'interprétation peuvent se poser avec les séries différenciées. L'autre avantage dans ce type de modèle est qu'on pourra isoler chacune des composantes d'intérêt pour l'étudier à part, à l'inverse de l'approche de Box & Jenkins où on ne s'intéresse qu'à la modélisation de la partie stationnaire de la série, après élimination des autres composantes par différenciation.

Les modèles à composantes inobservables se mettent facilement sous forme espace-état. On s'épargne ainsi les problèmes de stationnarité (au sens faible) et de racine unitaire qui se posent préalablement à l'estimation d'un modèle ARMA. En particulier, les résultats du filtre de Kalman restent valides en présence de séries non-stationnaires. Par ailleurs, ce cadre permet également de relâcher l'hypothèse d'une distribution gaussienne pour les bruits. D'autre part, l'estimation optimale prend en compte l'information disponible à partir de la date initiale t = 0, alors que les estimateurs optimaux ARMA prennent en compte l'information à partir de t = $-\infty$. Enfin, les coefficients du modèle peuvent évoluer et ne sont pas obligatoirement considérés comme invariants au cours de la période d'estimation.

Dans le modèle proposé par Harvey (1989), une série Y_t est décomposée de manière additive en une tendance, un cycle et une composante résiduelle.

$$Y_t = \mu_t + \psi_t + \varepsilon_t, \qquad t = 1, \ldots, T$$

Chacune de ces composantes est stochastique et elles sont supposées mutuellement non corrélées entre elles. Cette décomposition de base a été modifiée pour nous donner le modèle suivant :

$$Y_t = \mu_t + \gamma_t + \psi_t + r_t + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=1}^{m} \theta_j X_{j,t} + \varepsilon_t \quad , t = 1, \ldots, T$$

avec ε_t i.i.d et qui suivent une loi normale $N(0, \sigma_\varepsilon^2)$.

Les termes μ_t, γ_t, ψ_t et r_t représentent respectivement la tendance, la saisonnalité, la cyclicité et le composant autorégressif d'ordre 1 pour corriger la cyclicité. Ces termes modélisent structurellement différents aspects des séries temporelles. Ils sont supposés statistiquement

indépendants les uns des autres et indépendants de ε_t, le terme d'erreur. Les termes Y_{t-i} et $X_{j,t}$ modélisent respectivement les composantes autorégressives de Y_t et les variables explicatives.

La tendance :

Le paramètre μ_t modélise la tendance naturelle des séries, en l'absence de saisonnalité, de cycles ou d'effets d'aucunes variables indépendantes. Dans la modélisation ARMA, on fait l'hypothèse que la tendance est déterministe. Cette contrainte, très restrictive, n'existe plus dans les modèles UCM, où l'on peut faire varier dans le temps la moyenne et la pente qui compose la tendance. Cette dernière devient stochastique.

La tendance linéaire déterministe est de la forme : $\mu_t = \alpha + \beta t$. Puisque μ_t peut être obtenu récursivement par : $\mu_t = \mu_{t-1} + \beta$, avec $\mu_0 = \alpha$, la continuité peut être préservée en introduisant des termes stochastiques comme suit :

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim \text{i.i.d. } N(0, \sigma_\eta^2)$$

$$\beta_t = \beta_{t-1} + \xi_t, \qquad \xi_t \sim \text{i.i.d. } N(0, \sigma_\xi^2)$$

où η_t et ξ_t sont des bruits blancs mutuellement non corrélés. Le rôle du paramètre η_t est de permettre à la tendance de changer de niveau, de varier à la hausse ou à la baisse. Le paramètre ξ_t permet quant à lui à la tendance de varier sa pente. Plus les variances de ces deux paramètres sont élevées et plus les mouvements stochastiques (aléatoires) sont importants dans la tendance. La tendance ainsi modélisée est dite LLT (locally linear time trend) ou « tendance localement linéaire dans le temps ».

Si $\sigma_\xi^2 = 0$, on obtient une tendance linéaire avec une pente fixe.

Si $\sigma_\eta^2 = 0$, le modèle résultant a, en général, une tendance plus lissée.

Si $\sigma_\eta^2 = \sigma_\xi^2 = 0$, alors on retombe sur une tendance linéaire déterministe.

Si la pente de la tendance reste approximativement constante tout au long de la vie de série, sans aucune tendance à la hausse ou à la baisse, on a un modèle de « marche aléatoire » ou RW (Random-Walk) : $\mu_t = \mu_{t-1} + \eta_t, \qquad \eta_t \sim \text{i.i.d. } N(0, \sigma_\eta^2)$,

Le cycle :

Considérons le paramètre ψ_t comme une fonction cyclique du temps, avec une fréquence λ,

$0 < \lambda < \pi$. La période du cycle est de $2\pi/\lambda$.

Un cycle peut être exprimé comme un mélange de fonctions sinus et cosinus dépendant de 2 paramètres α et β : $\psi_t = \alpha \cos(\lambda t) + \beta \sin(\lambda t)$. Si l'échelle de temps est continue (T est un intervalle de IR), ψ_t est une fonction périodique, d'amplitude $(\alpha^2 + \beta^2)^{1/2}$ et de phase $\tan^{-1}(\beta/\alpha)$.

Comme la tendance linéaire, le cycle peut être exprimé de manière récursive :

$$\begin{pmatrix} \psi_t \\ \psi_t^* \end{pmatrix} = \begin{pmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{pmatrix} \begin{pmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{pmatrix}$$

avec comme conditions initiales : $\psi_0 = \alpha$ et $\psi_0^* = \beta$. Notons que ψ_0 et ψ_0^* vérifient la relation : $\psi_t^2 + \psi_t^{*2} = \alpha^2 + \beta^2$, $\forall$ t.

La généralisation stochastique du cycle ψ_t peut être obtenue en ajoutant un bruit aléatoire à la récursion et en introduisant un facteur d'amortissement ρ dit « damping factor », pour augmenter la flexibilité du modèle. ρ est un paramètre de lissage : plus il est proche de 0, plus le cycle est irrégulier; plus il est proche de 1 et plus le cycle est régulier.

Le modèle qu'on obtient pour la composante cyclique est le suivant :

$$\begin{pmatrix} \psi_t \\ \psi_t^* \end{pmatrix} = \rho \begin{pmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{pmatrix} \begin{pmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{pmatrix} + \begin{pmatrix} \nu_t \\ \nu_t^* \end{pmatrix}$$

avec $0 \le \rho \le 1$ et les perturbations ν_t et ν_t^* qui sont indépendantes et suivent une loi $N(0, \sigma_\nu^2)$. Le cycle stochastique qui en ressort a une période fixée mais une amplitude et une phase qui varie dans le temps. Le modèle est stationnaire si $\rho<1$; et si λ est égal à 0 ou π, il se réduit à un processus autorégressif de premier ordre.

Les cycles, à l'état pur, ne sont pas souvent utilisés dans la pratique. Cependant, ils sont très utiles comme composantes pour des modèles périodiques très complexes.

Le terme autorégressif :

On introduit un terme autorégressif d'ordre 1 dans le modèle pour le cas particulier où le cycle a une fréquence λ égale à 0 ou π. La modélisation à part de ce cas aide à l'interprétation et à l'estimation des paramètres. Le composant autorégressif r_t est modélisé comme suit : $r_t = \rho r_{t-1} + \nu_t, \nu_t \sim \text{i.i.d. } N(0, \sigma_\nu^2)$ avec $-1 \le \rho \le 1$.

On peut voir ce composant comme : $\psi_t = \rho(\cos\lambda\,\psi_{t-1} + \sin\lambda\,\psi^*_{t-1}) + \nu_t = \rho\psi_{t-1} + \nu_t$ quand λ = 0. Dans ce cas, le deuxième terme en ψ^*_t n'est pas nécessaire.

La saisonnalité :

Les fluctuations saisonnières sont une source commune de variation des séries temporelles. Ces fluctuations apparaissent à cause des changements réguliers de saisons ou d'autres évènements réguliers. Les effets saisonniers sont considérés comme des corrections à la tendance générale de la série à cause des variations saisonnières. Ces effets, sommés sur une période saisonnière entière, s'annulent. Par conséquent, la composante saisonnière γ_t est modélisée comme périodique stochastique de période entière *s* telle que la somme est toujours égale à zéro en moyenne.

On peut voir la saisonnalité comme un cas particulier de la composante cyclique : en l'absence de chocs sur le cycle ($\sigma^2_\nu = 0$) et avec un lissage maximal (ρ = 1), le cycle est une sinusoïde parfaite dont la période se déduit du paramètre λ par la formule 2π/λ.

Les deux différents modèles pour la composante saisonnière sont :

→ Le modèle de type « Dummy variable » :

L'équation suivante le décrit :

$$\sum_{i=0}^{s-1} \gamma_{t-i} = \omega_t, \qquad \omega_t \sim \text{i.i.d. N}(0, \sigma^2_\omega)$$

→ Le modèle avec la composante saisonnière de forme « trigonométrique » :

Dans ce modèle, γ_t est modélisée comme une somme de cycles de différentes fréquences. Le modèle est le suivant :

$$\gamma_t = \sum_{j=1}^{[s/2]} \lambda_{j,t}$$

avec [s/2] = s/2 si s est pair et [s/2] = (s-1)/2 si s est impair. Les cycles $\gamma_{j,t}$ ont des fréquences $\lambda_j = 2\pi j/s$ et sont spécifiés par la l'équation matricielle :

$$\begin{pmatrix} \gamma_{j,t} \\ \gamma^*_{j,t} \end{pmatrix} = \rho \begin{pmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{pmatrix} \begin{pmatrix} \gamma_{j,t-1} \\ \gamma^*_{j,t-1} \end{pmatrix} + \begin{pmatrix} \omega_{j,t} \\ \omega^*_{j,t} \end{pmatrix}$$

où les perturbations $\omega_{j,t}$ et $\omega^*_{j,t}$ sont supposés indépendantes et, à j fixé, $\omega_{j,t}$ et $\omega^*_{j,t}$ suivent une N $(0, \sigma^2_\omega)$. Si la période s est paire, l'équation avec $\lambda^*_{s/2,t}$ n'est pas nécessaire et $\lambda_{s/2,t}$ est donné par : $\lambda_{s/2,t} = -\lambda_{s/2,t} + \omega_{s/2,t}$.

Si $\sigma^2_\omega = 0$ alors la composante saisonnière est constante ou déterministe.

Les cycles $\gamma_{j,t}$ sont appelés des harmoniques. Si la saisonnalité est déterminée, la décomposition des effets saisonniers dans ces harmoniques est identique à la décomposition de Fourier. La caractéristique principale de la saisonnalité est que sa période (c'est le temps nécessaire pour effectuer un cycle entier) est connue.

Les termes de régression :

Les termes de régression $\sum_{j=1}^{m} \beta_j x_{j,t}$ et $\sum_{i=1}^{p} \varphi_i y_{t-i}$ apportent une flexibilité supplémentaire au modèle. Des retards, différences et autres transformations peuvent être appliquées aux variables.

Les estimations des paramètres que nous venons de définir sont effectuées avec le filtre de Kalman. On reformule le modèle à composantes inobservables précédent comme un modèle espace-état. Nous présenterons ici la reformulation du modèle de base de Harvey, sans toutes les variables explicatives. En effet, il faudrait une équation d'état supplémentaire pour chaque variable explicative rajoutée dans le modèle.

L'équation d'observation qu'on obtient avec le modèle $Y_t = \mu_t + \psi_t + \varepsilon_t$, modèle de type « tendance – cycle » est la suivante :

$$Y_t = (1\ 0\ 1\ 0)\ Z_t + \varepsilon_t$$

où Z_t est le vecteur d'état défini par $Z_t = {}^t(\mu_t, \beta_t, \psi_t, \psi^*_t$ **Erreur ! Signet non défini.**) qui suit *l'équation d'état* :

$$Z_t = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \rho\cos\lambda & \rho\sin\lambda \\ 0 & 0 & -\rho\sin\lambda & \rho\cos\lambda \end{pmatrix} Z_{t-1} + \begin{pmatrix} \eta_t \\ \xi_t \\ \nu_t \\ \nu^*_t \end{pmatrix} \text{ avec } \begin{pmatrix} \eta_t \\ \xi_t \\ \nu_t \\ \nu^*_t \end{pmatrix} \sim N\left(0, Diag(\sigma^2_\eta, \sigma^2_\xi, \sigma^2_\nu, \sigma^2_\nu)\right)$$

L'équation d'observation qu'on obtient avec le modèle $Y_t = \mu_t + \psi_t + \theta_t X_{j,t} + \varepsilon_t$, modèle avec une seule variable explicative, une tendance et un cycle est la suivante :

$$Y_t = (1\ 0\ 1\ 0)\ Z_t + \theta_t X_{j,t} + \varepsilon_t$$

où Z_t est le vecteur d'état défini par $Z_t = {}^t(\mu_t, \beta_t, \psi_t, \psi_t^*$ **Erreur ! Signet non défini.**) qui suit *l'équation d'état* :

$$Z_t = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \rho\cos\lambda & \rho\sin\lambda \\ 0 & 0 & -\rho\sin\lambda & \rho\cos\lambda \end{pmatrix} Z_{t-1} + \begin{pmatrix} \eta_t \\ \xi_t \\ \nu_t \\ \nu_t^* \end{pmatrix} \text{ avec } \begin{pmatrix} \eta_t \\ \xi_t \\ \nu_t \\ \nu_t^* \end{pmatrix} \sim N\left(0, Diag(\sigma_\eta^2, \sigma_\xi^2, \sigma_\nu^2, \sigma_\nu^2)\right)$$

Le coefficient de régression se réécrit : $\theta_t = \theta_{t-1} + \tau_t$ avec $\tau_t \sim N(0, \sigma_\tau^2)$. τ_t est généralement fixé à 0 pour établir une relation stable entre Y_t et X_t (c.-à-d. avoir $\theta_t = \theta_{t-1}$ = constante).

Les valeurs retardées de Y_t peuvent être introduites dans le modèle structurel et donnent l'équation suivante : $Y_t = \varphi_1 Y_{t-1} + ... + \varphi_r Y_{t-r} + (1\ 0\ 1\ 0)\ Z_t + \theta_t X_{j,t} + \varepsilon_t$, t = r+1, …, T.

Nous procédons de la même manière pour réécrire des modèles plus complexes, en rajoutant autant d'équations d'état que de variables explicatives dans le modèle.

L'estimation des variables d'état s'effectue avec le filtre de Kalman, puis celle des paramètres avec l'algorithme EM. L'algorithme EM est un algorithme itératif utilisé pour calculer les Estimateurs du Maximum de Vraisemblance (EMV) des paramètres d'un modèle espace-état. On trouvera plus de détails sur le fonctionnement de l'algorithme ainsi que sur la théorie du filtre de Kalman en annexe 7.

Le processus général d'estimation est le suivant : le vecteur d'initialisation (pour les paramètres libres) est fixé. La vraisemblance est alors maximisée avec l'algorithme EM et on obtient les valeurs estimées avec leur écart-type. Les variances des composantes sont estimées à 0 et la vraisemblance est concentrée par rapport à celles-ci. Ensuite un filtre de Kalman fournit les innovations ainsi que leur variance. Un lisseur appliqué aux différents bruits permet in fine de générer les différentes composantes estimées.

III.5. Régression linéaire multiple

III.5.1. Principes de base de la régression linéaire

La régression linéaire se classe parmi les méthodes d'analyses qui traitent des données quantitatives. C'est une méthode d'investigation sur données d'observation, ou d'expérimentation, où l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.

C'est la méthode la plus utilisée en traitement de données pour deux raisons majeures :

- c'est une méthode ancienne,
- c'est l'outil de base de la plupart des modélisations plus sophistiquées comme la régression logistique, le modèle linéaire généralisé, entre autres.

La régression linéaire multiple est une généralisation du modèle linéaire classique. La forme générale du modèle s'exprime comme suit :

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$, i = 1, …, n

où : Y_i est la variable expliquée à caractère aléatoire ;

i est l'indice de l'observation courante ;

β_0, β_1, …, β_k sont les paramètres du modèle, que nous estimons à l'aide des données par la méthode des moindres carrés (MCO) ou OLS (Ordinary Least Squares) ;

X_{i1}, X_{i2}, …, X_{ik} sont les variables explicatives du modèle, à caractère non aléatoire ;

ε_i est l'erreur aléatoire ;

Le critère des moindres carrés correspond à la minimisation de la somme des carrés des écarts (SC Erreur en français, SS Error en anglais) entre Y observée et Y estimée par l'équation de régression. Quand les erreurs sont normalement distribuées, l'estimateur des moindres carrés est l'estimateur le plus efficace pour la régression. Mais cet estimateur est très sensible à la présence de valeurs extrêmes ou aberrantes.

III.5.2. La régression en présence de valeurs extrêmes

Une visualisation graphique de la série avant toute analyse permet de les repérer.

On pourrait certes retirer de nos données toute observation située à une certaine distance de l'écart-type. Mais si ces points ne sont pas aberrants (c'est-à-dire le résultat d'une erreur de mesure ou de saisie des données), les retirer de nos observations n'est pas la solution, d'autant plus qu'on est dans un cadre temporel. On ne tiendrait pas compte de l'autocorrélation qui existe dans le cas des séries temporelles. D'un autre côté, ne rien faire n'est pas une solution.

a. Régression linéaire robuste

On peut être amené à traiter ces points atypiques avec des méthodes dites « robustes ». Les méthodes de régression robuste ne sont pas très utilisées, même si les logiciels statistiques courants (R, SAS etc.) arrivent à les incrémenter. Il est généralement admis que les méthodes de régression robustes doivent respecter les propriétés suivantes :

- → l'efficacité : le taux de convergence vers les vraies valeurs des paramètres par l'algorithme d'estimation ;
- → une forte valeur de rupture : la valeur de rupture mesure la proportion nécessaire d'observations « contaminées » dans le jeu de données pour changer l'estimation des paramètres.

Parmi les nombreux estimateurs existants, nous n'utiliserons que le MM-estimateur.

Les MM-estimateurs :

La classe des MM-estimateurs a été introduite par Yohai (1987) dans la régression linéaire. Pour mieux comprendre la MM-estimation, une explication de la M-estimation et de la S-estimation est nécessaire. Les compléments théoriques sont disponibles en annexe 6.

La M-estimation :

La M-estimation a été introduite par Huber, le M est mis pour «Maximum-Likelihood-Like». La méthode consiste à minimiser la somme d'une fonction des résidus, qui augmentent moins rapidement, et sont donc plus robustes plutôt que de minimiser la somme des carrés. La fonction des résidus doit être choisie de manière à limiter l'influence des valeurs atypiques. La résolution se fait par itération de l'algorithme IRLS (« iteratively reweighted least-squares ») ; l'idée est d'estimer les coefficients avec une régression par les moindres carrés dans laquelle on donne aux résidus faibles un poids plus fort qu'aux résidus élevés.

Cet estimateur est robuste pour les points atypiques dans la variable de réponse, mais pas pour ceux des variables explicatives. Quand il y a des outliers dans les variables explicatives, cette méthode n'a plus de réel avantage sur la méthode des MCO.

La S-estimation :

Les S-estimateurs pour la régression multiple univariée ont été introduits par Rousseeuw et Yohai(1984). Cette méthode tire son nom du paramètre de dispersion S qui apparait dans une équation qui doit y être résolue. Elle égalise la valeur attendue des résidus pondérés (normalisés par une valeur S) à la valeur attendue si les données étaient normalement distribuées. Comme avec les fonctions de poids dans la M-estimation, la fonction de poids n'est pas aussi fortement affectée par les valeurs élevées des résidus que dans les MCO. L'estimation trouve les paramètres β_i qui minimise S avec un algorithme itératif.

La MM-estimation :

Cette méthode est une combinaison des méthodes précédentes. L'intérêt de ces estimateurs est qu'ils conservent la robustesse et la résistance du S-estimateur, mais en gagnant l'efficacité du M-estimateur. L'estimation se fait selon les étapes suivantes :

- un paramètre d'échelle est estimé en utilisant une équation similaire à celle de la S-estimation, avec les résidus pondérés ;
- ce paramètre d'échelle est ensuite utilisé dans une M-estimation qui a une seconde fonction de poids.

b. Introduction de variables d'intervention

Une des solutions possibles consiste à considérer ces observations extrêmes comme des interventions. On conceptualise la donnée extrême comme une intervention unique, de type saut (voir partie III.3.Régression avec erreurs ARMA), et on l'introduit dans le modèle statistique avant de faire la régression.

Nous venons d'expliciter la théorie qui nous servira dans les modélisations que nous allons maintenant effectuer. Les méthodes décrites ci-dessus seront appliquées sur les différentes séries, en utilisant le logiciel SAS, avec les modules BASE, STAT et ETS ; Le but étant de déterminer le modèle adéquat pour chacune de nos séries étudiées.

IV. MISE EN APPLICATION SUR NOS DONNEES

IV.1. Réalisation des modèles de type ARMA

Avant de commencer la modélisation, nous avons testé la stationnarité de nos séries, pour conforter les observations faites dans la partie descriptive. En effet, les modèles ARMA ne sont applicables qu'à des séries stationnaires. Pour les séries qui n'étaient pas stationnaires (présence de saisonnalité et/ou de tendance), la méthode de la différenciation a été retenue.

Nous avons utilisé la procédure ARIMA du logiciel SAS pour effectuer cette modélisation. Nous détaillerons la méthode utilisée pour la série des fréquences mensuelles en RC corporelle ; pour les autres séries, les résultats seront résumés, et les analyses complémentaires disponibles en annexe 8.

Exemple de la série des fréquences mensuelles en RC corporelle :

L'analyse exploratoire nous révélait la saisonnalité de notre série, ainsi qu'une tendance à la baisse. Le tracé des fonctions d'autocorrélation et d'autocorrélation partielle nous confirmait ces impressions.

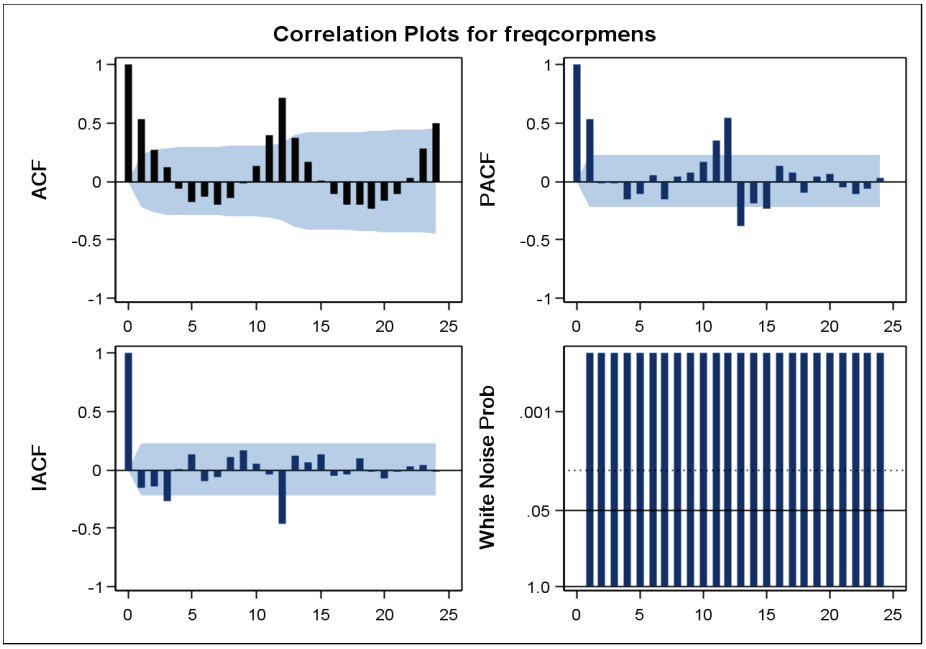


Figure 11 : Fonctions d'autocorrélation des fréquences mensuelles en RC corporelle

Nous avons constaté que la fonction d'autocorrélation était oscillatoire, avec un pic qui revenait tous les 12 mois (notre unité temporelle étant le mois). Nous avons alors appliqué à la série l'opérateur $\nabla_{12} = 1 - B^{12}$. A la suite de quoi, nous avons obtenu les graphes suivants :

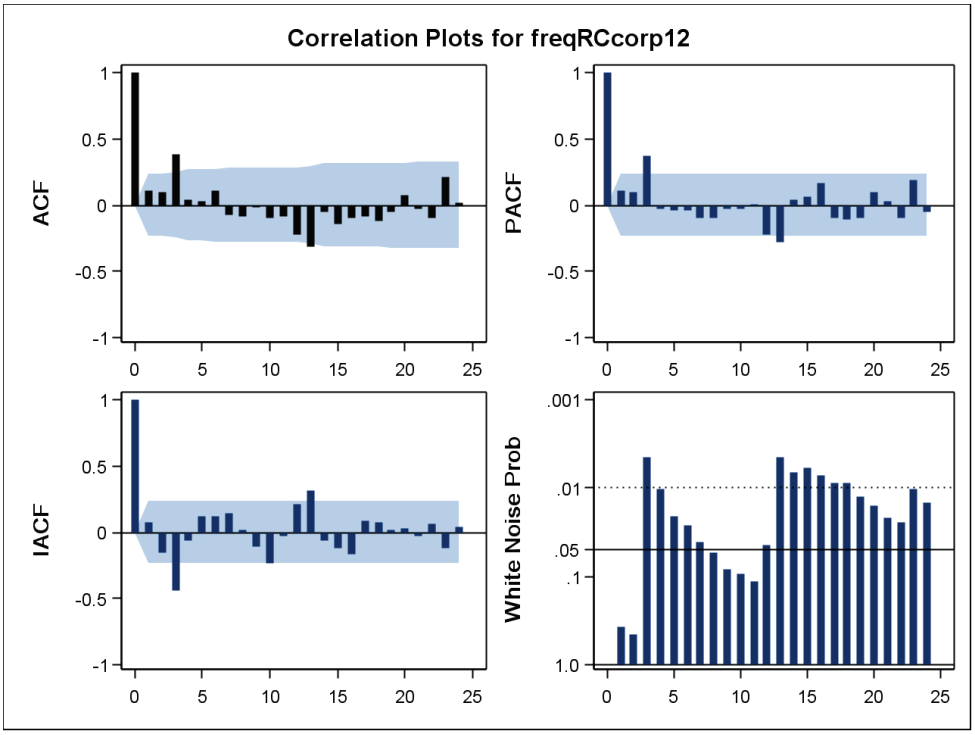


Figure 12 : Fonctions d'autocorrélation de la série désaisonnalisée des fréquences

La rapide décroissance de la fonction d'autocorrélation (ACF) vers 0 montre le gain de la différenciation. Le test du porte-manteau nous montre que nous n'avons pas un bruit blanc (graphique « White Noise Probability ») : les probabilités d'être supérieures à la statistique du Khi 2 sont trop faibles et ce dès le 3ème retard.

L'ACF ne nous montre pas la présence d'une tendance. Si la tendance persistait, on aurait eu une fonction d'autocorrélation empirique lentement décroissante, ce qui n'est pas ici le cas. L'ACF et le PACF (fonction d'autocorrélation partielle) décroissent très rapidement après 0, mais présentent un pic au retard 3. On décide toutefois de tester la présence ou non de racines unités, pour savoir si on doit différencier une fois de plus notre série.

Nous avions choisi un modèle sans constante, donc nous effectuons un test de Dickey-Fuller augmenté. Les résultats des tableaux numéro et numéro sont obtenus avec la macro STATIONARITY de Dominique Ladiray :

variable	test	Type	Stat	Lag0	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7	Lag8
freqRCcorp12	ADF	Zero Mean	Tau	-6,349	-4,007	-2,190	-2,021	-1,843	-1,707	-1,741	-1,727	-1,753
			Pr < Tau	0,000	0,000	0,028	0,042	0,063	0,083	0,077	0,080	0,076

variable	test	Type	Stationary	Lag0	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7	Lag8
freqRCcorp12	ADF	Zero Mean	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No

Tableau 3 : Résultats des tests de racine unité au seuil de 5 %

Nous avons conclu que notre série n'avait pas de racines unités. On l'a donc modélisée avec un AR(3). Les coefficients d'ordre 1 et 2 du processus AR, de même que la constante n'étaient pas significatifs. L'équation obtenue est : $(1 - 0{,}51\ B^3)$ fréquences mensuelles $_t = \varepsilon_t$. Le coefficient du processus autorégressif est bien de module inférieur à 1.

Un dernier regard sur les résidus nous a permis de valider notre modèle. Ils sont bien la réalisation d'un bruit blanc : vérification avec l'ACF, le PACF et le test du porte-manteau.

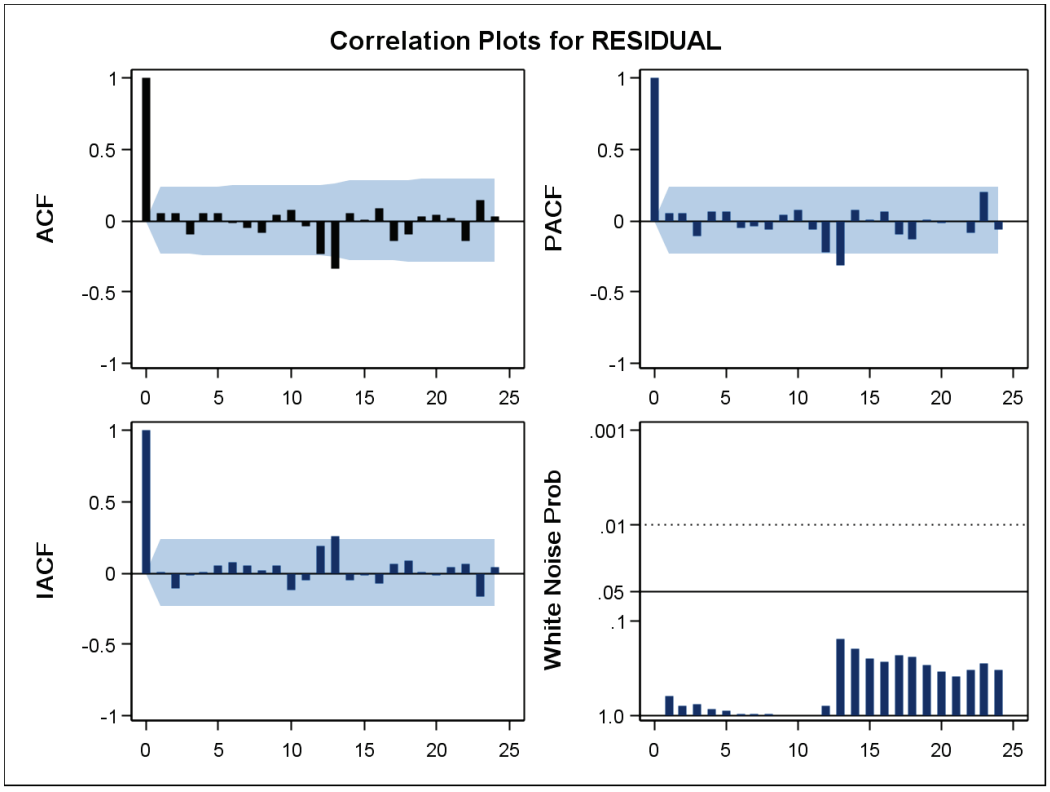


Figure 13 : Fonction d'autocorrélation des résidus du modèle

Nous remarquions un léger pic au retard d'ordre 13 ; mais en effectuant les tests à l'ordre 13, nous avons conclu que ce pic n'était pas significatif.

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	4.18	5	0.5236	0.108	0.109	-0.034	0.109	0.122	0.049
12	8.79	11	0.6408	0.017	-0.015	0.097	0.131	0.022	-0.159
18	18.53	17	0.3560	-0.244	0.108	0.069	0.148	-0.076	-0.047
24	24.96	23	0.3523	0.075	0.082	0.061	-0.098	0.173	0.068

Tableau 4 : Test de blancheur des résidus du modèle

Le tableau «Autocorrelation Check of Residuals» nous montre qu'aucune des Q-statistiques n'est statistiquement significative. En effet, la probabilité d'être supérieure à la statistique du Khi 2 est bien supérieure à 5%. Ce qui veut dire que notre modèle s'ajuste bien à nos données.

Le tableau ci-dessous résume les résultats que nous avons obtenus pour les séries de fréquences mensuelles :

Pour les séries de fréquences mensuelles :

Garanties	Saisonnalité	Tendance	Partie AR	Partie MA	Constante
RC Corporelle	$(1 - B^{12})$	NC *	$(1 - 0{,}51\ B^{3})$	NC	NC
RC matérielle	$(1 - B^{12})$	NC	$(1 - 0{,}51\ B^{3})$	NC	NC
Dommages	$(1 - B^{12})$	NC	NC	NC	NC
Bris de glace	$(1 - B^{12})$	NC	$(1 - 0{,}93\ B)$	$(1 - 0{,}78\ B)$	NC
Incendie	NC	NC	NC	NC	NC
Vol total	NC	$(1 - B)$	NC	$(1 - 0{,}80\ B)$	-0,0002
Vol partiel	$(1 - B^{12})$	$(1 - B)$	NC	$(1 - 0{,}79\ B)$	NC

* : Non concernée

Pour la RC matérielle, nous obtenons un modèle $SARIMA_{12}$ [(3,0,0)(0,0,0)].
Pour la garantie Dommages, une fois la saisonnalité retirée, la série devient un bruit blanc.
La série des fréquences en Bris de glace suit un modèle $SARIMA_{12}$ [(1,0,1)(0,0,0)].
La série des fréquences de sinistres en Incendie n'est pas modélisable par un modèle ARMA. Elle serait un bruit blanc selon l'ACF et le test du porte-manteau. Mais un doute subsiste car la représentation graphique nous montrait la présence de points extrêmes dans la série. Or ce type de points perturbe l'autocorrélogramme de la série.

La série des fréquences en Vol total suit un modèle ARIMA(0,1,1). Pour la série des fréquences en Vol partiel, le modèle retenu est un $SARIMA_{12}$ [(0,1,1)(0,0,0)].

Pour les séries de coûts de sinistres clos :

Garanties	Saisonnalité	Tendance	Partie AR	Partie MA
RC Corporelle	$(1 - B^{12})$	NC*	(1 – 0,52 B)	NC
RC matérielle	$(1 - B^{12})$	(1 – B)	(1 + 0,35 B)	NC
Dommages	$(1 - B^{12})$	(1 – B)	NC	NC
Bris de glace	NC	(1 – B)	NC	NC
Incendie	NC	NC	NC	NC
Vol total	NC	(1 – B)	NC	NC
Vol partiel	$(1 - B^{12})$	(1 – B)	NC	NC

* : Non concernée

La RC corporelle et la RC matérielle sont les seules garanties dont les coûts sont modélisables par un processus AR. Pour les autres séries de coûts, une fois la tendance et/ou la saisonnalité retirée, nous n'avons plus que des bruits blancs.

Pour les séries de coûts de sinistres déclarés :

N'ayant pas obtenu de résultats très concluants sur l'ensemble des séries de coûts de sinistres clos, nous nous intéressons à la modélisation des coûts de sinistres déclarés.

Les modèles obtenus sont résumés dans le tableau ci-dessous :

Garanties	Saisonnalité	Tendance	Partie AR	Partie MA
RC matérielle	$(1 - B^{12})$	(1 – B)	$1 + 0{,}32\ B^2$	$1 - 0{,}69\ B^{12}$
Dommages	$(1 - B^{12})$	(1 – B)	NC	NC
Bris de glace	$(1 - B^{12})$	(1 – B)	$1 + 0{,}43\ B^{12}$	NC
Incendie	$(1 - B^{12})$	NC	1 – 0, 54 B	NC
Vol total	$(1 - B^{12})$	(1 – B)	NC	1 – 0,73 B
Vol partiel	$(1 - B^{12})$	(1 – B)	$1 + 0{,}89\ B^{12}$	1 – 0,67 B

* : Non concerné

On constate que la saisonnalité apparait dans tous les modèles de coûts de sinistres déclarés. On obtient des modèles pour toutes les garanties sauf pour la garantie Dommages.

IV.2. Réalisation des modèles de régression avec erreurs ARMA

Notre point de départ a été l'étude des liens de corrélation établis dans la partie II.2.2. En effet, la corrélation n'implique pas la causalité, mais l'existence d'une relation de causalité entraine automatiquement la corrélation. Donc l'absence de corrélation implique l'absence de causalité.

Les résultats obtenus pour les différentes garanties avec les procédures ARIMA et AUTOREG en testant la pertinence de nos variables d'intervention sont résumés ci-dessous :

Pour les fréquences mensuelles :

Variables explicatives potentielles	RC corporelle	RC matérielle	Dommages	Bris de glace	Vol partiel	Vol total
Constante		-0,723	0,640	0,195	-0,025	0,052
Ensoleillement/100				0,048		
Parcours mensuel - autoroute interurbaine (21)				1,051		
Parcours mensuel - Autoroute et voie rapide urbaine (22)			-0,281			
Parcours mensuel - Route nationale interurbaine à caractéristiques autoroutière (23)				-1,118		-0,012
Vitesse de jour sur les RD		0,009				
Vitesse de jour sur les RN 2-2/100						0,018
Volume de Super	0,34				0,784	0,201
Volume de Gazole	0,06		1,549	2,142		
Volume de Super + Gazole		1,616	0,731			
AR (1)			0,309			
AR (3)	0,43	0,323			0,553	
AR (10)						-0,352
X1	- 0,023	-0,182	-0,089	-0,253		
X2	- 0,014	-0,112	-0,112		0,015	
X3	- 0,011	-0,109	-0,052		0,010	
X4	- 0,012	-0,120	-0,084		0,007	
X5	- 0,011	-0,078	-0,031		0,012	
X6		-0,026			0,012	
X7	- 0,008	-0,101	-0,079			

X8	- 0,009	-0,124	-0,143			
X9		-0,039	-0,016		0,010	
X10		-0,028			0,009	
X11					0,021	

En RC corporelle, le modèle que nous obtenons reflète encore la saisonnalité, vu le nombre important de variables d'intervention qu'il intègre. C'est également le cas pour les séries de fréquences en RC matérielle, en Dommages et en Vol partiel : la prise en compte de la saisonnalité s'impose. Sur la garantie Bris de glace, le mois de janvier apparait comme atypique mais la saisonnalité est absorbée par les variables explicatives. Sur la garantie Vol total, le retard d'ordre 10 ressort dans le modèle.

Les tests sur les résidus de ces modèles ainsi que les graphiques des modèles obtenus sont disponibles respectivement en annexe 9.

Les conclusions qui ressortent de cette modélisation sont les suivantes :

Pour les séries de fréquences, les variables qui influent sur nos séries sont celles liées à la livraison de carburant, au nombre de kilomètres parcourus ainsi qu'à la vitesse sur certains axes routiers. Ce sont des modèles assez cohérents avec la fréquence de sinistres automobile.

Pour les séries de coûts de sinistres clos, qui devenaient des bruits blancs après le retrait de la tendance et de la saisonnalité avec la modélisation ARMA, le modèle de régression avec erreurs ARMA ne nous apporte aucune information supplémentaire. Nous ne pouvons toujours pas avoir de manière explicite l'expression de la tendance et/ou de la saisonnalité.

Pour les coûts de sinistres déclarés, nous n'avons pas obtenu de modèles avec des variables explicatives qui soient pertinentes.

Pour continuer notre étude en tenant compte de l'aspect temporel de nos séries, de la présence de tendance et de saisonnalité, il nous faut les décomposer. C'est ce qui nous a amené à faire de la modélisation des composantes inobservables : ce sont les modèles UCM (Unobserved Components Models).

IV.3. Réalisation des modèles de type UCM

Nous avions dans la partie Modélisation ARMA, détecté des séries pour lesquelles, une fois supprimées la saisonnalité et la tendance, nous n'avions plus que des bruits blancs. Pour ces cas, nous voulons néanmoins obtenir des résultats, qui nous permettront de faire nos prévisions. En décomposant les séries avec un modèle UCM, nous arriverons à obtenir les variances des composantes de la série et nous pourrons également essayer de les expliquer avec d'autres variables.

Nous avons testé deux types de modèles : les modèles saturés (sans variables explicatives) et les modèles non-saturés (avec variables explicatives). Dans la suite, nous vous présenterons les modèles obtenus selon les garanties. L'étude des résidus pour ces modèles a été mise en annexe 10.

Avec la procédure UCM de SAS, les paramètres des modèles que nous obtenons sont une somme de composantes et des paramètres de type régressif. L'estimation que nous effectuons est celle des variances des termes de perturbation, des coefficients d'amortissement et des fréquences des cycles ainsi que des coefficients de régression des termes de régression. Ces différents paramètres sont estimés par l'algorithme EM et le filtre de Kalman.

La logique que nous adoptons est celle d'une « stepwise », avec une élimination progressive des paramètres non significatifs du modèle.

- Pour les paramètres de type régressif, si les p-values des coefficients ne sont pas significatives, les variables doivent être retirées, une variable à la fois (on retire la variable la moins significative) jusqu'à ce que toutes les variables restantes soient significatives.
- Pour les composantes, la création d'un modèle parcimonieux se fait en deux étapes. La première étape consiste à déterminer si la composante varie dans le temps. La seconde est de déterminer si elle contribue au modèle. L'hypothèse de départ est de considérer que les composantes du modèle sont à la fois stochastiques et significatives. La composante peut être significative, sans varier dans le temps (être déterministe).

Nous présenterons tout d'abord les modèles sans variables explicatives (ou modèles saturés) et les modèles avec variables explicatives (ou modèles non-saturés).

Modèles saisonniers saturés :

Séries de fréquences mensuelles

Pour la garantie RC corporelle :

Nous n'avons réalisé qu'un modèle saisonnier saturé pour la modélisation de cette garantie. La série présente une saisonnalité (de période 12) et une tendance que nous devrons retrouver dans le modèle final.

L'ajustement des différents paramètres s'est effectué comme suit :

Nous avons commencé par modéliser un modèle complet, avec toutes ses composantes (tendance, cycle, bruit blanc, saisonnalité de période 12) stochastiques. Le modèle nous a proposé un cycle de période 26, avec une composante non significative et qui ne correspondait pas à l'observation de notre série. On a donc retiré la composante cyclique et on a relancé le modèle. Cette fois, la pente n'était plus significative dans le modèle. On l'a retirée et relancé la procédure. La composante Bruit blanc n'étant toujours pas significative, on l'a également retirée du modèle en égalisant sa variance à 0. Le modèle final ne contient plus que le niveau et la saisonnalité et converge en 5 itérations.

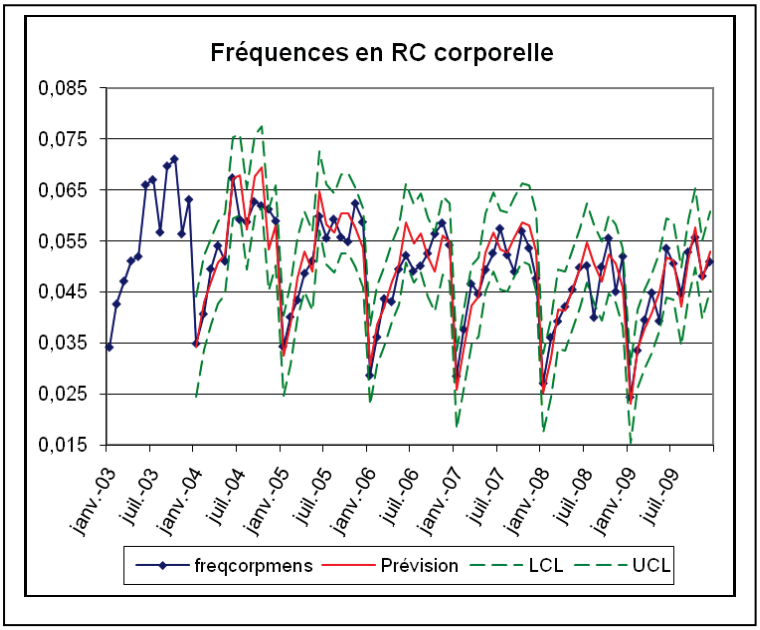
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	1192.90	<.0001
Season	11	572.89	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00001580
Root Mean Squared Error	0.00397
Mean Absolute Percentage Error	6.72909
Maximum Percent Error	13.32106
R-Square	0.81291
Adjusted R-Square	0.81024
Random Walk R-Square	0.83701
Amemiya's Adjusted R-Square	0.80222
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 72	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	0.00000125	5.5164E-7	2.26	0.0239
Season	**Error Variance**	0.00000517	1.24885E-6	4.14	<.0001

La saisonnalité est une composante déterministe (on a forcé sa variance à 0).

Le modèle est de la forme : Fréquences $_t = \mu_t + \gamma_t$ avec une moyenne à 0,04484.



L'hypothèse de blancheur des résidus ainsi que celle de normalité ne sont pas rejetées.

Pour la garantie RC matérielle :

De même que pour la garantie RC corporelle, nous n'avons trouvé parmi nos variables explicatives aucune d'entre elles qui expliquaient les fréquences en RC matérielle. Nous n'avons donc modélisé que le modèle saturé. Le retard d'ordre 3 qui ressortait dans la modélisation ARMA n'a pu être éliminé par les autres composantes. Nous l'avons réintroduit dans le modèle qui est :

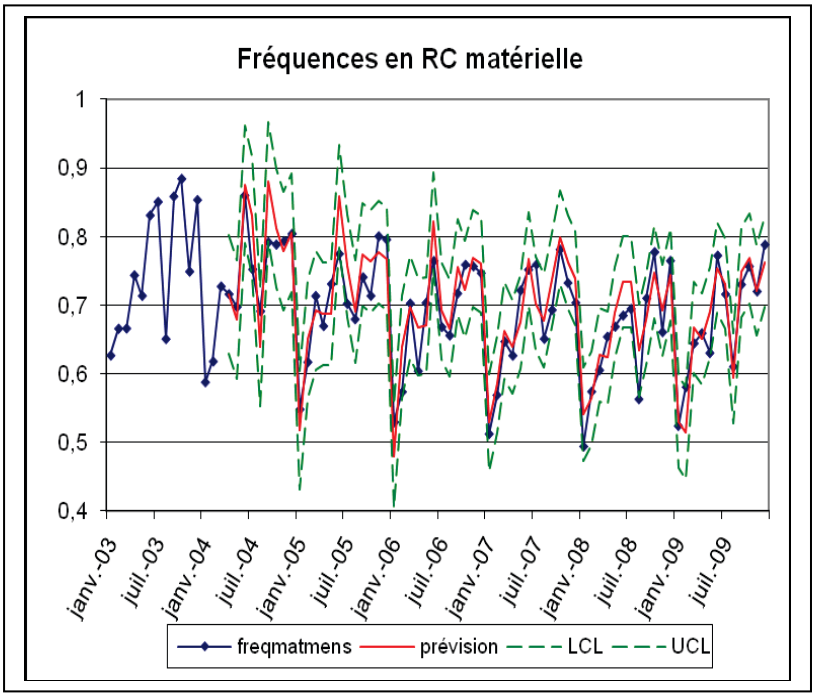
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	3.60	0.0578
Level	1	2890.81	<.0001
Season	11	782.16	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00138
Root Mean Squared Error	0.03708
Mean Absolute Percentage Error	4.55512
Maximum Percent Error	11.60682
R-Square	0.78789
Adjusted R-Square	0.78473
Random Walk R-Square	0.91614
Amemiya's Adjusted R-Square	0.77523
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 69	

Le retard d'ordre 3 est significatif, avec un coefficient égal à 0,73, la tendance et la saisonnalité sont déterministes. L'hypothèse de blancheur des résidus ainsi que celle de normalité ne sont pas rejetées. Le modèle est de la forme :

Fréquences $_t = \mu_t + \gamma_t + 0{,}73$ Fréquences $_{t-3} + \varepsilon_t$.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00096881	0.0001649	5.87	<.0001
DepLag	**Phi_1**	0.73018	0.07609	9.60	<.0001



Pour la garantie Dommages :

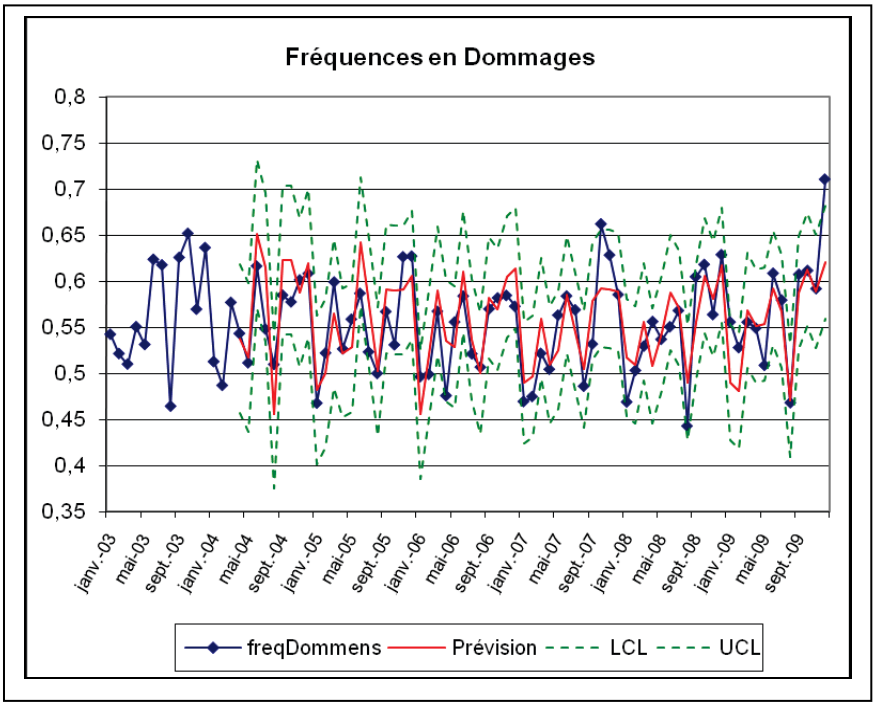
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	49.09	<.0001
Level	1	10242.8	<.0001
Season	11	219.52	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00116
Root Mean Squared Error	0.03408
Mean Absolute Percentage Error	4.96772
Maximum Percent Error	12.66171
R-Square	0.56506
Adjusted R-Square	0.55857
Random Walk R-Square	0.86445
Amemiya's Adjusted R-Square	0.53910
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 69	

Le retard d'ordre 3 est significatif, avec un coefficient égal à 0,41, la tendance et la saisonnalité sont déterministes. L'hypothèse de blancheur des résidus et celle de normalité ne sont pas rejetées.

Le modèle obtenu est : Fréquences $_t = \mu_t + \gamma_t + 0{,}41$ Fréquences $_{t-3} + \varepsilon_t$.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00084729	0.0001443	5.87	<.0001
DepLag	**Phi_1**	0.41004	0.11434	3.59	0.0003



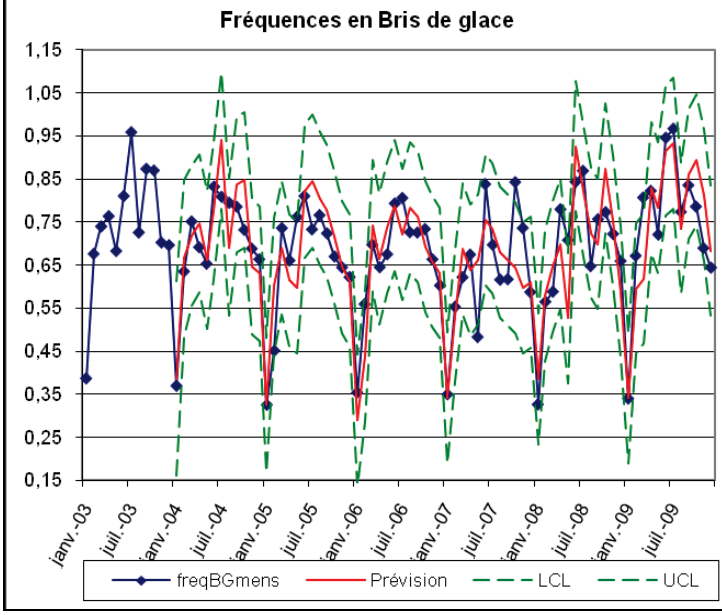
Pour la garantie Bris de glace :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	560.57	<.0001
Season	11	277.93	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00611
Root Mean Squared Error	0.07815
Mean Absolute Percentage Error	9.13543
Maximum Percent Error	25.63320
R-Square	0.68362
Adjusted R-Square	0.67910
Random Walk R-Square	0.73477
Amemiya's Adjusted R-Square	0.66554
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 72	

La tendance et la saisonnalité sont stochastiques. La composante Bruit blanc a été supprimée du modèle, l'aléa restant porté par chacune des composantes retenues. Le modèle obtenu est le suivant : Fréquences $_t = \mu_t + \gamma_t$, avec une moyenne à 0,6897.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	0.00085259	0.0003629	2.35	0.0188
Season	**Error Variance**	0.00151	0.0004412	3.43	0.0006



L'hypothèse de blancheur des résidus ainsi que celle de normalité ne sont pas rejetées.

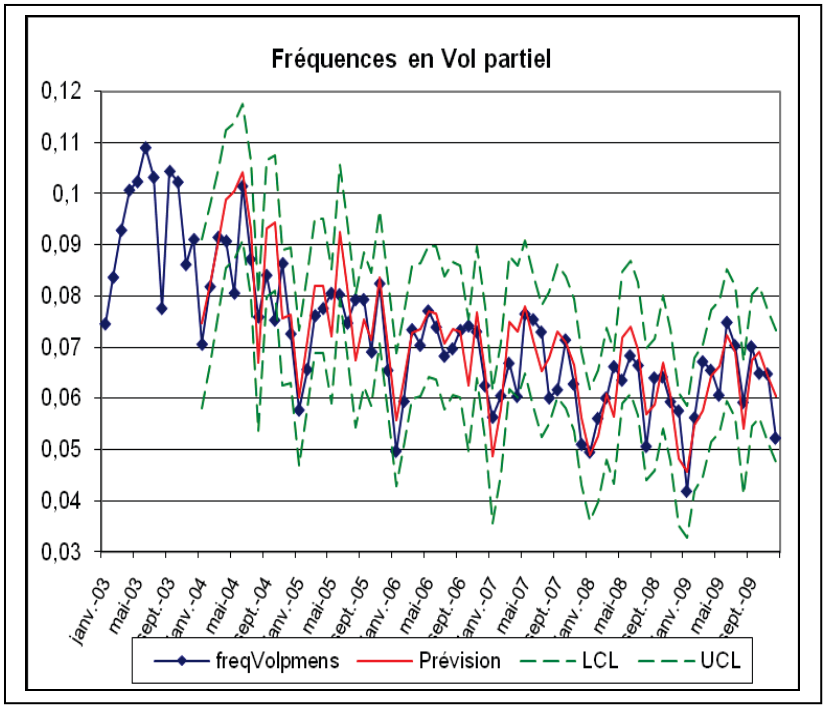
Pour la garantie Vol partiel :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	789.31	<.0001
Season	11	266.57	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00004463
Root Mean Squared Error	0.00668
Mean Absolute Percentage Error	7.72223
Maximum Percent Error	16.39220
R-Square	0.62628
Adjusted R-Square	0.62094
Random Walk R-Square	0.70855
Amemiya's Adjusted R-Square	0.60492
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 72	

La tendance et la saisonnalité sont stochastiques. La composante Bruit blanc a été supprimée du modèle. Le modèle obtenu est le suivant : Fréquences $_t = \mu_t + \gamma_t$.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	0.00000356	1.40373E-6	2.54	0.0112
Season	**Error Variance**	0.00001434	3.25446E-6	4.41	<.0001



L'hypothèse de blancheur des résidus et celle de normalité ne sont pas rejetées.

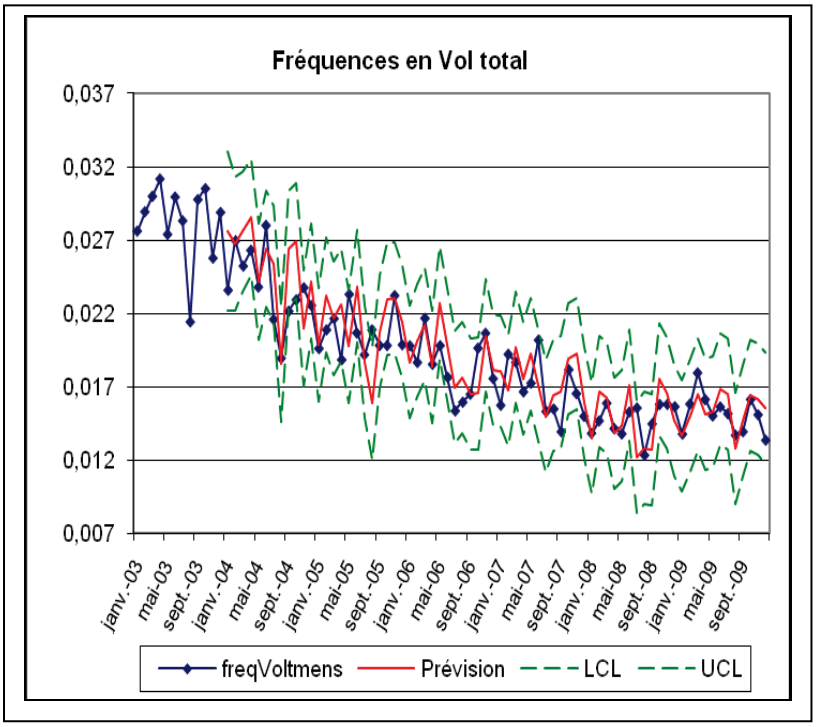
Pour la garantie Vol total :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	393.28	<.0001
Season	11	50.13	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00000397
Root Mean Squared Error	0.00199
Mean Absolute Percentage Error	8.47477
Maximum Percent Error	23.99005
R-Square	0.69259
Adjusted R-Square	0.68820
Random Walk R-Square	0.68697
Amemiya's Adjusted R-Square	0.67503
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 72	

La tendance et la saisonnalité sont stochastiques. La composante Bruit blanc a été supprimée du modèle. Le modèle obtenu est le suivant : Fréquences $_t = \mu_t + \gamma_t$.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	4.667944E-7	1.53885E-7	3.03	0.0024
Season	**Error Variance**	0.00000102	2.34259E-7	4.35	<.0001



L'hypothèse de blancheur des résidus et celle de normalité ne sont pas rejetées.

Pour la garantie Incendie :

La série des fréquences en Incendie ne présentait pas de saisonnalité. On remarquait des points extrêmes et une tendance quasi-linéaire. Avec la modélisation ARMA, nous n'arrivions pas à capter cette tendance. En effet, la présence de ces points augmente la variance de la série, réduit le pouvoir explicatif du modèle, réduit la puissance des tests et altère le graphe de l'autocorrélogramme de la série. Avec la modélisation UCM, nous ne nous attarderons pas plus sur cette série. Nous sommes toutefois parvenus à estimer une tendance linéaire égale à une fréquence de 0,0094, pour un écart-type de 0,0004.

Séries de coûts de sinistres clos

Pour les séries de coûts clos, nous avons refait la même analyse. Les résultats obtenus sont détaillés ci-dessous :

Pour la garantie RC corporelle :

La branche RC étant à déroulement long, nous n'avons modélisé que les coûts de sinistres clos. Le modèle obtenu est : Fréquences $_t$ = – 0,18 Fréquences $_{t-2}$ + μ_t + γ_t.

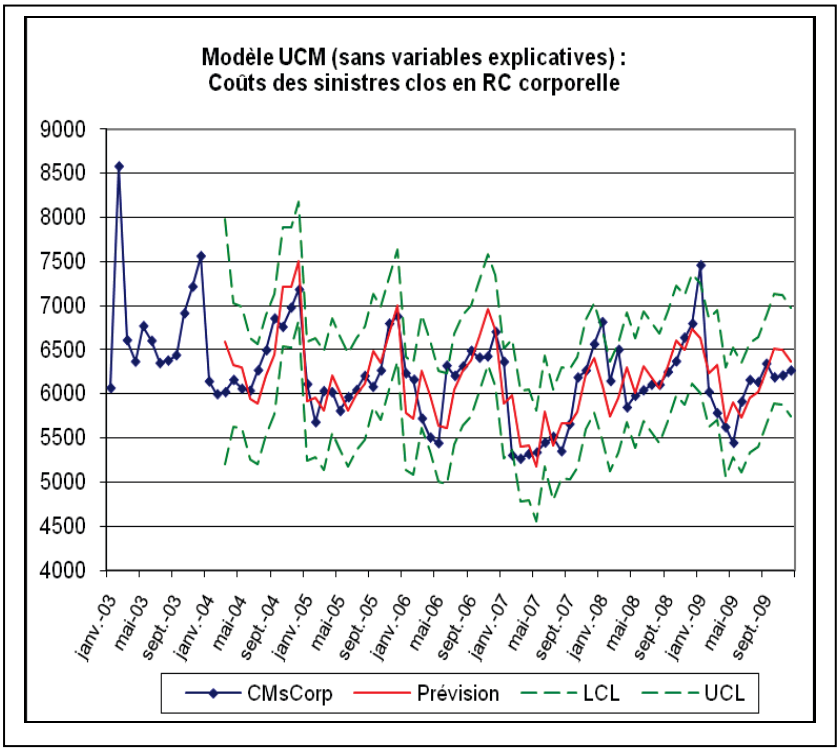
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	2575.24	<.0001
Season	11	100.77	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	108007
Root Mean Squared Error	328.64360
Mean Absolute Percentage Error	4.31377
Maximum Percent Error	11.22005
R-Square	0.49381
Adjusted R-Square	0.47870
Random Walk R-Square	0.83371
Amemiya's Adjusted R-Square	0.44848
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 70	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	40003	11034.2	3.63	0.0003
Season	**Error Variance**	9627.74456	3984.6	2.42	0.0157
DepLag	**Phi_1**	-0.17677	0.08293	-2.13	0.0330

On a introduit dans le modèle le retard d'ordre 2. On constate que le R carré ajusté du modèle est faible (0,48), même si l'erreur moyenne de notre modèle est de 4% pour des coûts qui sont de l'ordre de 6 000€.

La tendance et la saisonnalité sont stochastiques. La composante Bruit blanc a été supprimée du modèle. L'hypothèse de blancheur des résidus n'est pas rejetée.



Pour la garantie RC matérielle :

Nous avons dû introduire un retard d'ordre 2 dans notre modèle. Les résultats des différents tests et l'adéquation de notre modèle à la série sont résumés ci-dessous :

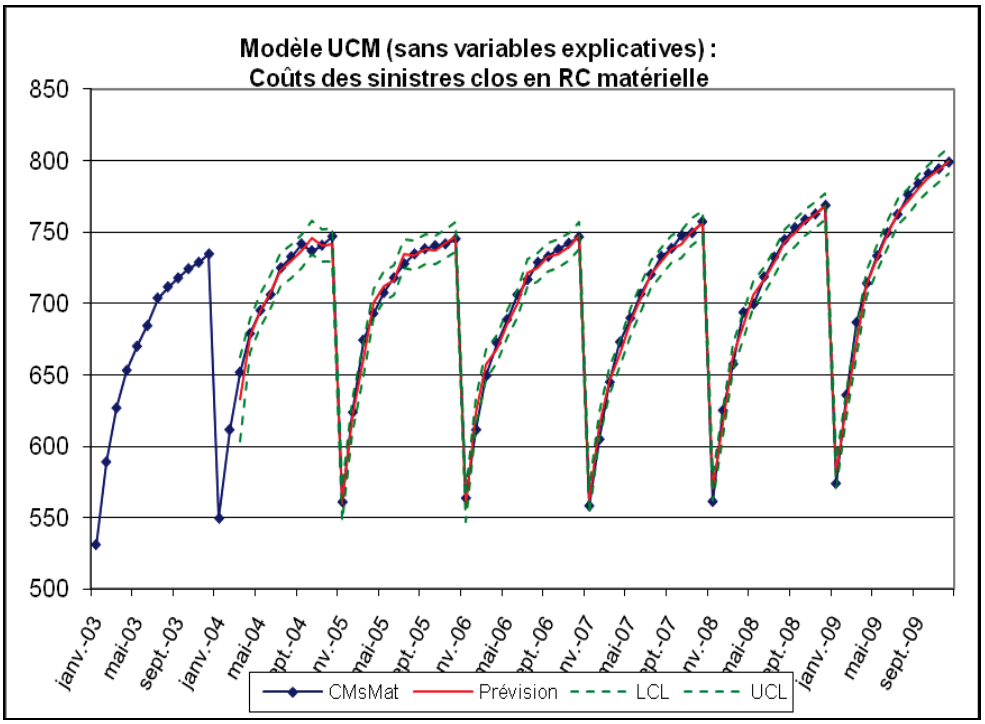
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	87229.9	<.0001
Season	11	15970.8	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	30.02264
Root Mean Squared Error	5.47929
Mean Absolute Percentage Error	0.58781
Maximum Percent Error	2.88815
R-Square	0.99114
Adjusted R-Square	0.99101
Random Walk R-Square	0.99666
Amemiya's Adjusted R-Square	0.99062
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 70	

La tendance est stochastique, mais la saisonnalité est déterministe. La composante Bruit blanc a été supprimée du modèle. L'hypothèse de blancheur des résidus n'est pas rejetée.

Le modèle obtenu est le suivant : $\text{Coût}_t = \mu_t + \gamma_t + 0{,}33\ \text{Coût}_{t-2}$.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	18.84495	3.18538	5.92	<.0001
DepLag	**Phi_1**	0.33057	0.11241	2.94	0.0033



Pour la garantie Dommages :

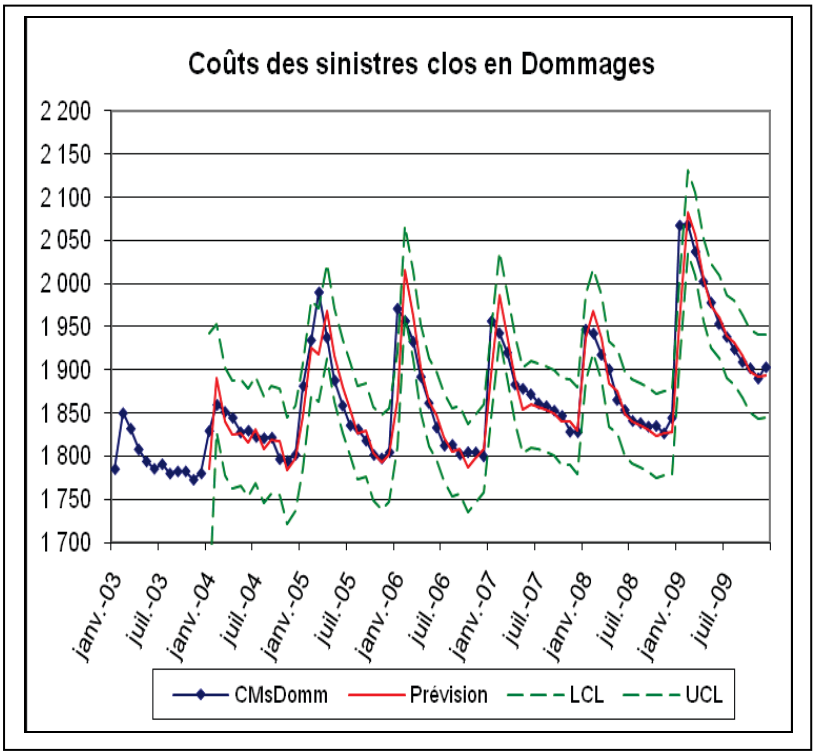
Nous avons modélisé les coûts clos en Dommages, car nous obtenions de meilleurs résultats que sur les coûts de sinistres déclarés. Le modèle obtenu est résumé ci-dessous :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Level	1	49518.1	<.0001
Season	11	230.15	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	730.89514
Root Mean Squared Error	27.03507
Mean Absolute Percentage Error	0.87807
Maximum Percent Error	5.39707
R-Square	0.83137
Adjusted R-Square	0.83137
Random Walk R-Square	0.98474
Amemiya's Adjusted R-Square	0.82662
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 72	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	529.25166	88.20861	6.00	<.0001

La tendance est stochastique, mais la saisonnalité est déterministe. La composante Bruit blanc a été supprimée du modèle. Le modèle obtenu est le suivant : Coût $_t = \mu_t + \gamma_t$. L'hypothèse de blancheur des résidus n'est pas rejetée.



Séries de coûts de sinistres déclarés

Excepté en RC corporelle et en Dommages, les séries de coûts déclarés des autres garanties ont été modélisées. Les résultats obtenus sont présentés dans cette partie.

Pour la garantie RC matérielle :

Le modèle retenu est résumé ci-dessous :

Coût $_t = \mu_t + \gamma_t - 0{,}28$ Coût $_{t-2}$.

Une précision s'impose ici sur les coûts de sinistres déclarés : on constate qu'on a des coûts négatifs dans notre série. Cela s'explique par la convention IDA (Indemnisation Directe de l'Assuré) : il s'agit d'une convention établie entre assureurs qui organise le règlement (rapide) des accidents automobiles.

L'assureur de responsabilité civile règle directement les dommages subis par le véhicule de son propre assuré, au lieu et place de l'assureur de l'auteur responsable. Un barème forfaitaire de responsabilité, élaboré à partir du Code de la route et de la jurisprudence, permet au vu du constat amiable signé par les différentes parties, de définir les responsabilités.

En fin d'année, l'assureur de l'auteur responsable rembourse, selon la convention, celui de la partie adverse. Ce sont ces montants perçus qui expliquent les coûts moyens de sinistres déclarés en négatif en RC matérielle.

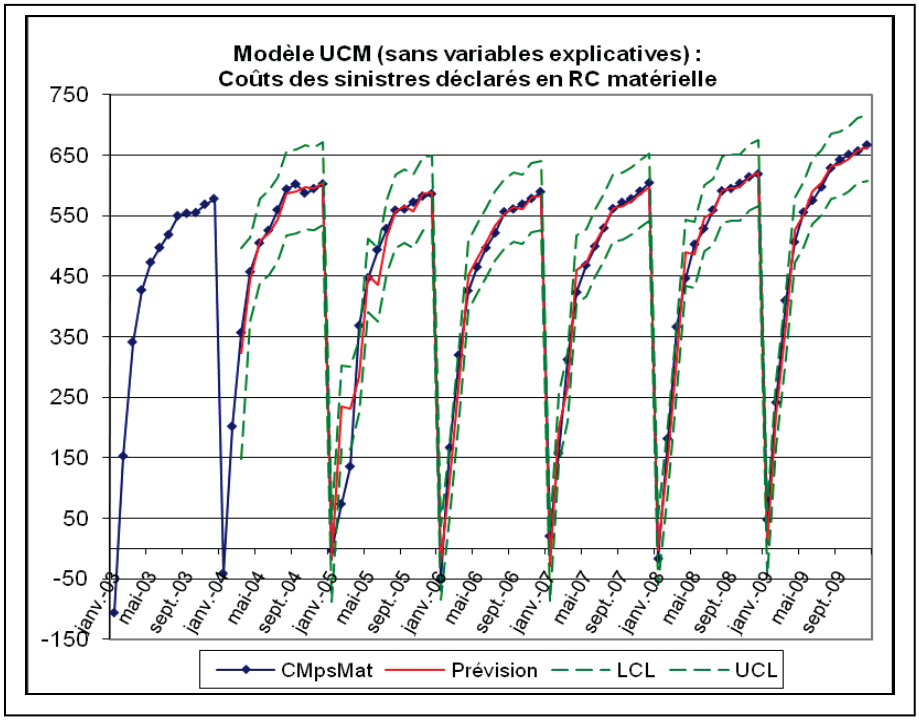
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	4388.80	<.0001
Season	11	4072.76	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	1034.84413
Root Mean Squared Error	32.16899
Mean Absolute Percentage Error	22.08802
Maximum Percent Error	243.57622
R-Square	0.96939
Adjusted R-Square	0.96894
Random Walk R-Square	0.96895
Amemiya's Adjusted R-Square	0.96759
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 70	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	658.98066	111.38806	5.92	<.0001
DepLag	**Phi_1**	-0.28283	0.10643	-2.66	0.0079

La tendance est stochastique, la saisonnalité déterministe (de période 12) et un retard d'ordre 2 a été introduit dans le modèle. L'hypothèse de blancheur des résidus n'est pas rejetée.

On constate un niveau d'erreur très élevé qui est dû aux coûts moyens observés durant l'année 2005 (mois de février-mars-avril et juin-juillet). Mais le modèle reste très bon sur le reste des mois comme on peut le constater avec le R carré ajusté et le graphique ci-dessous :



Pour la garantie Bris de glace :

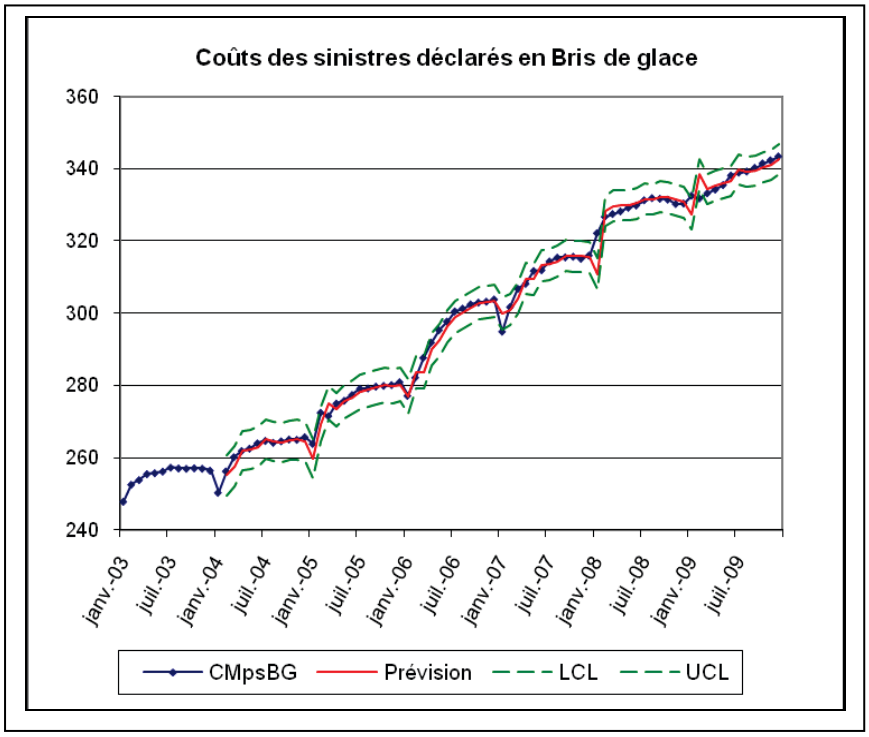
Le modèle retenu est résumé ci-dessous :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	207731	<.0001
Slope	1	26.23	<.0001
Season	11	53.92	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	5.08513
Root Mean Squared Error	2.25502
Mean Absolute Percentage Error	0.46324
Maximum Percent Error	3.51676
R-Square	0.99304
Adjusted R-Square	0.99304
Random Walk R-Square	0.99441
Amemiya's Adjusted R-Square	0.99284
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	3.90504	0.65541	5.96	<.0001

Le modèle est de la forme : Coût $_t = \mu_t + \beta_t + \gamma_t$.



L'hypothèse de blancheur des résidus n'est pas rejetée.

Pour la garantie Vol partiel :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	0.75	0.3855
Level	1	26283.7	<.0001
Slope	1	802.17	<.0001
Cycle	2	10.07	0.0065
Season	11	315.71	<.0001

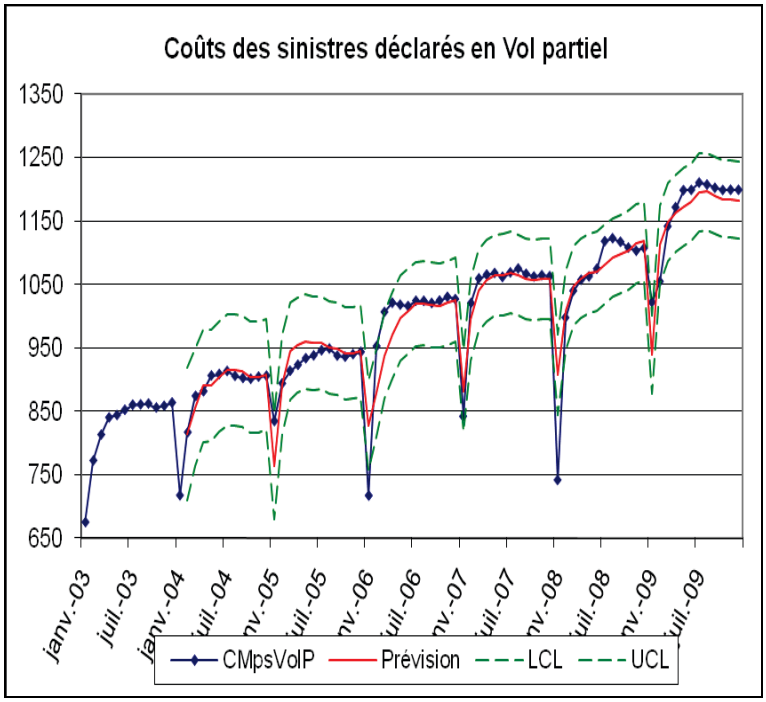
Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	1140.81172
Root Mean Squared Error	33.77590
Mean Absolute Percentage Error	2.07451
Maximum Percent Error	8.43903
R-Square	0.90953
Adjusted R-Square	0.90548
Random Walk R-Square	0.92132
Amemiya's Adjusted R-Square	0.89873
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

On a introduit une composante cyclique de période 31 dans le modèle.

Le modèle est de la forme : Coût $_t = \mu_t + \gamma_t + \psi_t + \varepsilon_t$.

L'hypothèse de blancheur des résidus n'est pas rejetée.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	702.36004	126.94115	5.53	<.0001
Cycle	**Damping Factor**	0.98739	0.01620	60.94	<.0001
Cycle	**Period**	31.36340	3.29905	9.51	<.0001
Cycle	**Error Variance**	8.18491	6.78186	1.21	0.2275

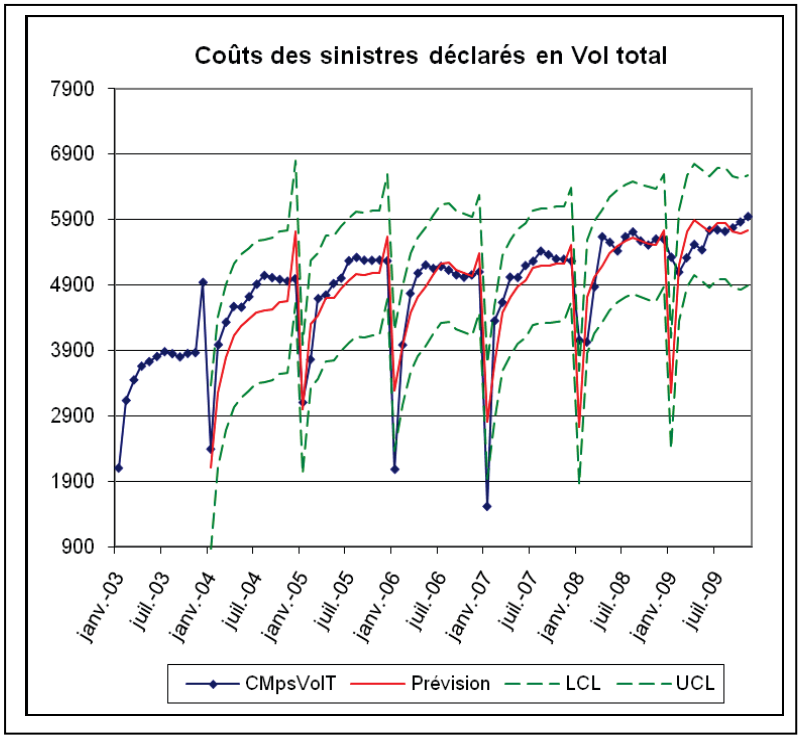


Pour la garantie Vol total :

Analyse de significativité des composantes (basée sur l'état final)			
Composante	**DDL**	**Khi-2**	**Pr > Khi-2**
Level	1	2417.10	<.0001
Season	11	190.26	<.0001

Statistiques d'ajustement basées sur les résidus	
Mean Squared Error	250617
Root Mean Squared Error	500.61685
Mean Absolute Percentage Error	7.75421
Maximum Percent Error	39.95081
R-Square	0.56693
Adjusted R-Square	0.56065
Random Walk R-Square	0.61710
Amemiya's Adjusted R-Square	0.54182
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Valeur estimée**	**Erreur type approchée**	**Valeur du test t**	**Approx. de Pr > |t|**
Level	**Error Variance**	186346	31275.6	5.96	<.0001
DepLag	**Phi_1**	-0.47590	0.10449	-4.55	<.0001



On a introduit un retard d'ordre 1 dans le modèle. La tendance est stochastique, la saisonnalité déterministe. Le modèle est de la forme : Coût $_t = \mu_t + \gamma_t - 0{,}48$ Coût $_{t-1}$.
L'hypothèse de blancheur des résidus n'est pas rejetée.

Modèles saisonniers non saturés :

Nous avons cherché à améliorer nos modèles avec l'intégration de variables explicatives.

Séries de fréquences mensuelles

Pour les séries de fréquences, les garanties qui nous ont donné des résultats avec les variables explicatives sont la RC corporelle et le Bris de glace.

Pour la garantie RC corporelle :

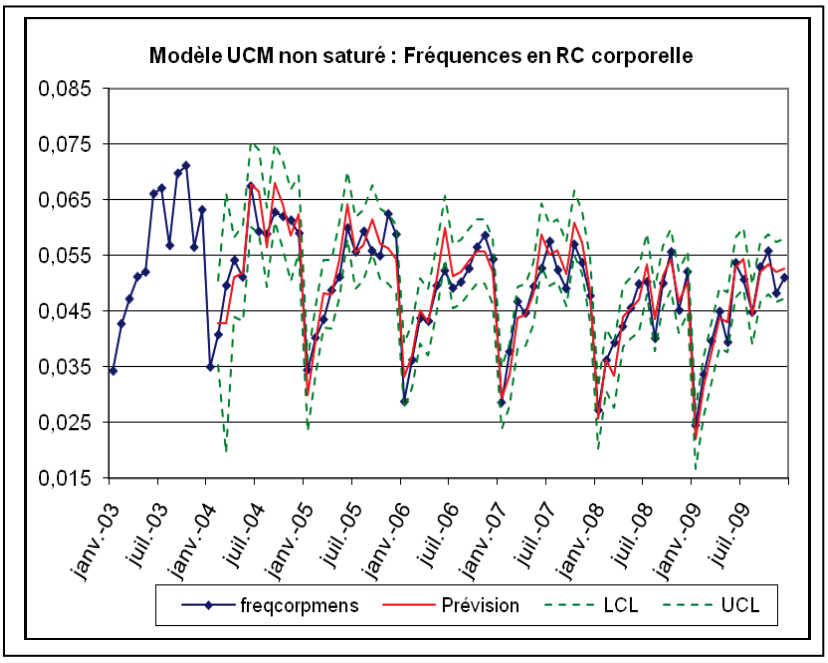
Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00000420	1.00913E-6	4.16	<.0001
Level	**Error Variance**	8.337288E-7	5.12852E-7	1.63	0.1040
Super_gazole_m3	**Coefficient**	0.15969	0.01915	8.34	<.0001

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	0.44	0.5052
Level	1	8.13	0.0044
Season	11	592.38	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00000999
Root Mean Squared Error	0.00316
Mean Absolute Percentage Error	5.38019
Maximum Percent Error	14.86975
R-Square	0.87948
Adjusted R-Square	0.87773
Random Walk R-Square	0.90392
Amemiya's Adjusted R-Square	0.87249
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

Le modèle s'écrit : Fréquence $_t = \mu_t + \gamma_t + 0{,}16$ Volume de Super et Gazole + ε_t.

La tendance est stochastique, la saisonnalité déterministe.



Pour la garantie Bris de glace :

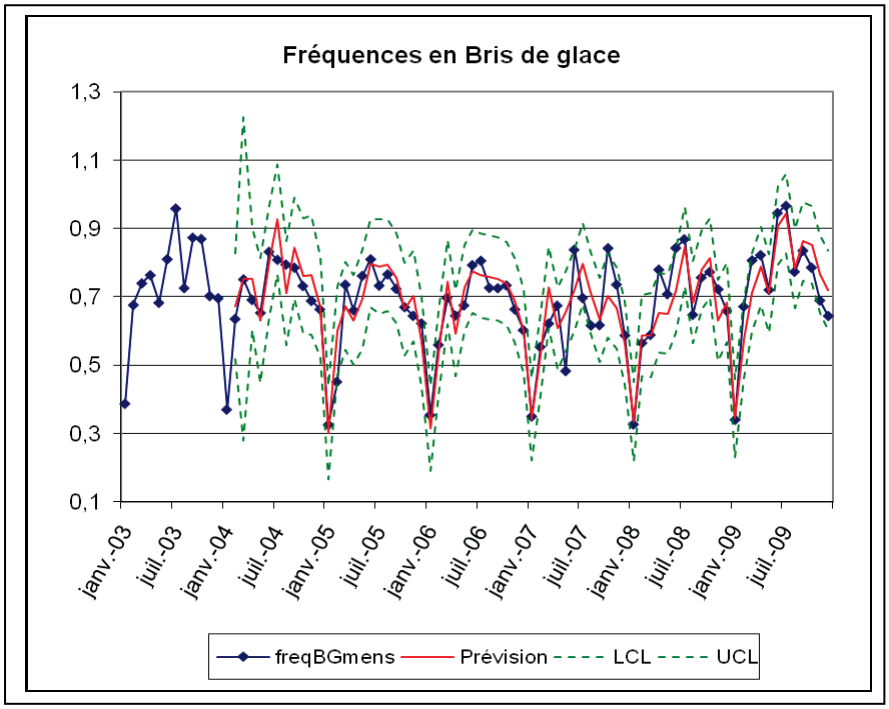
Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00207	0.0004418	4.69	<.0001
Level	**Error Variance**	0.00026919	0.0001630	1.65	0.0988
Super_gazole_m3	**Coefficient**	2.00369	0.41834	4.79	<.0001

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	2.40	0.1214
Level	1	0.55	0.4603
Season	11	247.83	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00402
Root Mean Squared Error	0.06344
Mean Absolute Percentage Error	7.37183
Maximum Percent Error	16.51200
R-Square	0.77853
Adjusted R-Square	0.77532
Random Walk R-Square	0.84590
Amemiya's Adjusted R-Square	0.76569
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

Le modèle s'écrit : Fréquence $_t = \mu_t + \gamma_t + 2{,}0$ Volume de Super et Gazole + ε_t.

La tendance est stochastique, la saisonnalité déterministe de période 12.



Séries de coûts de sinistres clos

Pour les coûts de sinistres clos, les garanties pour lesquelles nous avons pu introduire des variables explicatives sont la RC matérielle et le Bris de glace.

Pour la garantie RC matérielle :

Le modèle retenu intègre (en plus des composantes comme le niveau, la saisonnalité et la partie bruit blanc) l'indice des prix à la consommation (noté IPC) des pièces détachées et accessoires. Il s'écrit :

Coût $_t = \mu_t + \gamma_t + 3{,}36$ IPC pièces et accessoires $+ \varepsilon_t$.

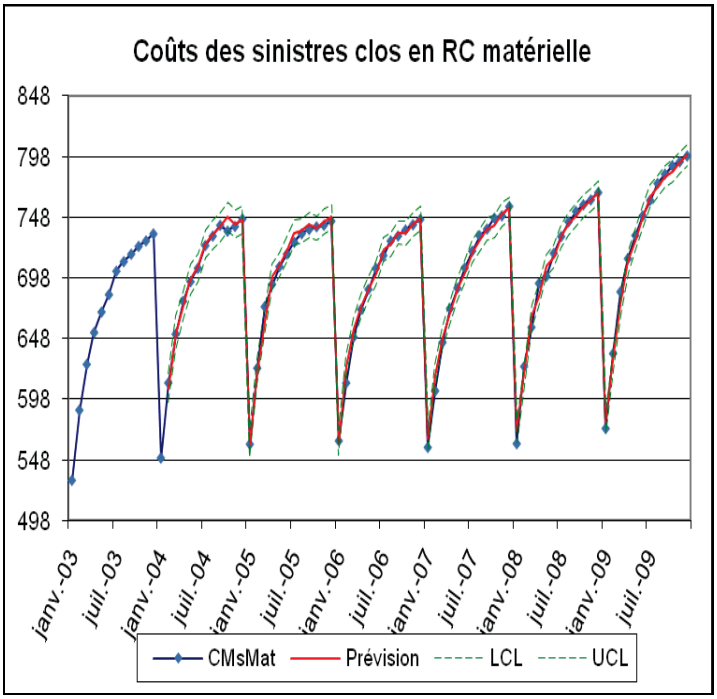
La tendance est stochastique et la saisonnalité déterministe de période 12.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Valeur estimée**	**Erreur type approchée**	**Valeur du test t**	**Approx. de Pr > |t|**
Level	**Error Variance**	18.67240	3.13391	5.96	<.0001
ipc_piece_access	**Coefficient**	3.36298	1.12881	2.98	0.0029

Analyse de significativité des composantes (basée sur l'état final)			
Composante	**DDL**	**Khi-2**	**Pr > Khi-2**
Irregular	0	.	.
Level	1	4.25	0.0393
Season	11	10757.9	<.0001

Statistiques d'ajustement basées sur les résidus	
Mean Squared Error	25.41437
Root Mean Squared Error	5.04127
Mean Absolute Percentage Error	0.54246
Maximum Percent Error	1.77907
R-Square	0.99267
Adjusted R-Square	0.99267
Random Walk R-Square	0.99176
Amemiya's Adjusted R-Square	0.99246
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

L'hypothèse de blancheur des résidus n'est pas rejetée.



Pour la garantie Bris de glace :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	0.00	0.9723
Season	11	35.21	0.0002

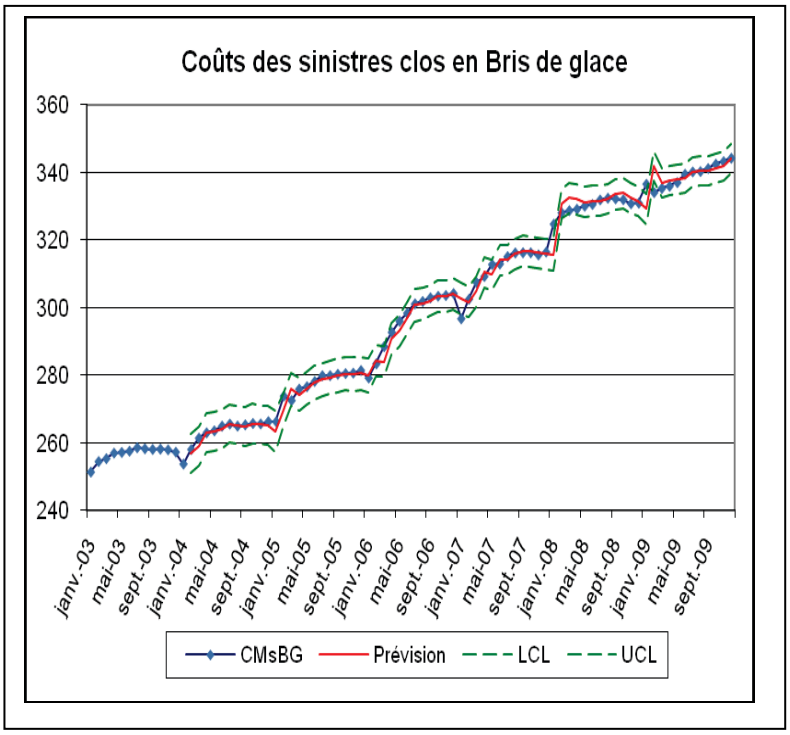
Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	5.53569
Root Mean Squared Error	2.35281
Mean Absolute Percentage Error	0.49109
Maximum Percent Error	2.76569
R-Square	0.99250
Adjusted R-Square	0.99250
Random Walk R-Square	0.99399
Amemiya's Adjusted R-Square	0.99229
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 71	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	4.19517	0.70410	5.96	<.0001
ipc_repar_veh_perso	**Coefficient**	2.24820	0.43466	5.17	<.0001

Le modèle s'écrit :

Coût $_t = \mu_t + \gamma_t + 2{,}24$ IPC réparation de véhicules personnels et accessoires.

La tendance est stochastique, la saisonnalité déterministe de période 12.



Les tests ne rejettent pas l'hypothèse de blancheur des résidus.

IV.4. Réalisation des modèles de régression linéaire multiple

Notre étude tenait à conserver une approche temporelle sur les différentes séries. Or l'analyse des fréquences mensuelles en Incendie nous a montré que la série ne présentait pas de comportement saisonnier, mais plutôt une tendance linéaire, avec des valeurs extrêmes. La régression simple nous a semblé intéressante pour effectuer cette modélisation.

Fréquences mensuelles en Incendie :

Le graphique de la série est représenté ci-dessous :

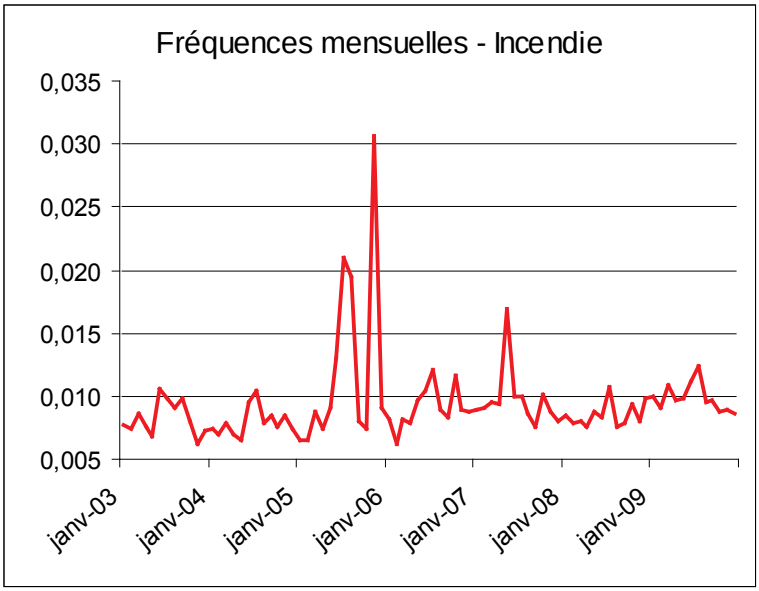


Figure 14 : Graphique des fréquences mensuelles en Incendie

On remarque des pics en juillet - août 2005 (une fête du 14 juillet particulièrement agitée), en novembre 2005 (émeutes en Ile de France et dans les régions) et en mai 2007 (élections présidentielles). Ces évènements ne sont pas traités par les variables externes dont nous disposons. Ce ne sont pas des valeurs aberrantes, mais plutôt des points atypiques, des points extrêmes dans la série.

Nous avons décidé de tenir compte de ces points extrêmes qui apparaissaient sur le graphique de la série c'est-à-dire ne pas les exclure de notre modélisation. Pour cela, nous avons appliqué les méthodes de régression robustes définies dans la partie « Régression multiple ».

Avec la méthode de la MM estimation :

La procédure ROBUSTREG de SAS nous permet justement de détecter et de faire des régressions en présence de valeurs extrêmes, avec une MM-estimation. Nous avons spécifié dans les options que c'est avec le S-estimateur que nous souhaitions construire notre MM-estimateur (par défaut, SAS initialise la procédure avec le LTS-estimateur).

Le modèle obtenu est le suivant :

Incendie = $0{,}0059 \times$ Parcours mensuel sur les autoroutes et voies rapides urbaines - $0{,}0067 \times$ Parcours mensuel sur route nationale interurbaine à caractéristiques autoroutières – $0{,}0093 \times$ Parcours mensuel sur les autres routes nationales + $0{,}0018 \times$ Parcours mensuel total sur autoroutes + $0{,}0001 \times$ Températures + ε_t

Avec un R^2 égal à 0,81, le modèle a détecté 5 outliers (voir annexe 11). Le graphique ci-dessous nous donne une idée de la précision de notre modèle.

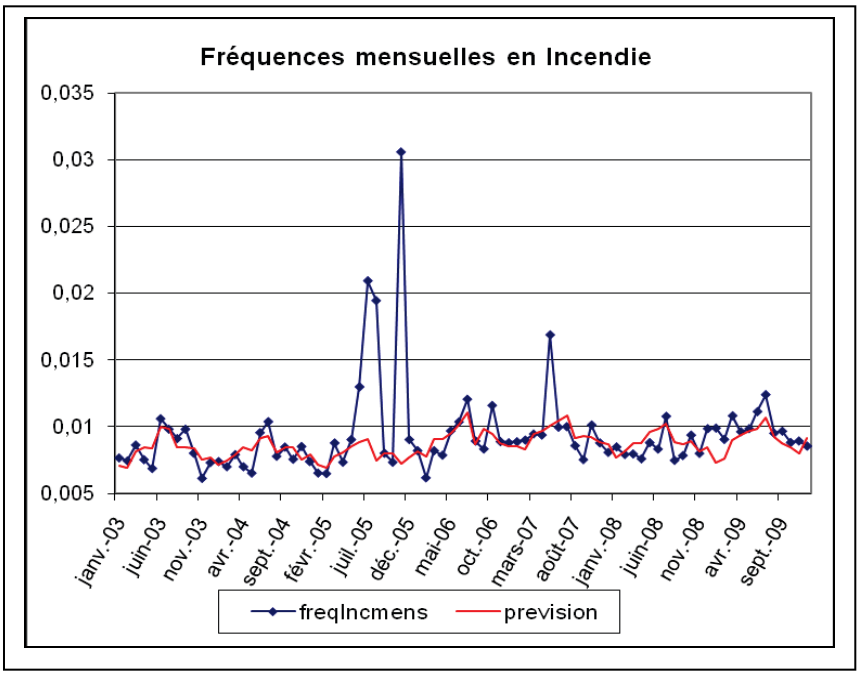


Figure 15 : Prévisions des fréquences en Incendie avec la MM-estimation

On constate que le modèle obtenu a associé aux valeurs extrêmes le MM-estimateur correspondant. Les résidus que nous obtenons sont bien un bruit blanc, même si ils ne sont pas gaussiens (voir annexe 11).

Avec la méthode d'introduction de variables d'intervention ponctuelles :

Nous avons appliqué la seconde méthode définie dans la partie III.5.2.b avec l'introduction des variables d'intervention de type sauts.

Pour cela, les variables que nous avons créées sont :

- → pulse1 qui vaut 1 en juillet et août 2005, 0 sinon
- → pulse2 qui vaut 1 en novembre 2005, 0 sinon
- → pulse3 qui vaut 1 en mai 2007, 0 sinon.

En les intégrant dans le modèle, avec la procédure AUTOREG de SAS, nous avons obtenu le modèle suivant avec un R^2 de 0,98 :

Incendie = 0,0018 Parcours mensuel sur autoroutes et voies rapides urbaines + 0,0003 Parcours moyen mensuel sur le total des autoroutes + 0,0108 pulse1 + 0,0225 pulse2 + 0,0076 pulse3.

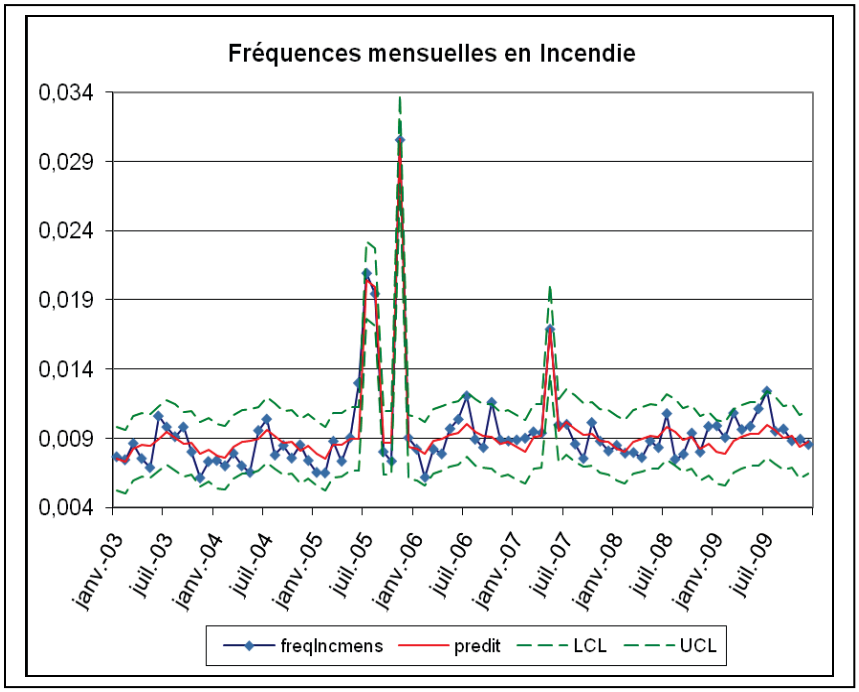


Figure 16 : Prévisions des fréquences en incendie avec les variables d'intervention

Les hypothèses de blancheur et de normalité des résidus ne sont pas rejetées.

Ce modèle nous permet d'augmenter notre R^2, en conservant les valeurs extrêmes. Néanmoins, en utilisant ces variables d'intervention, nous sous-entendons que les évènements qui ont provoqués ces valeurs extrêmes sont des évènements rares, qui ne se reproduiront pas. C'est une hypothèse assez forte.

V. VALIDATION DES DIFFERENTS MODELES

Pour tester la stabilité dans le temps des différents modèles obtenus, nous avons réalisé une autre modélisation sur les données tronquées à fin 2008. Ensuite, nous avons effectué avec ces nouveaux modèles une prévision de l'année 2009.

V.1. Cas des modèles ARMA

Pour les séries de fréquences mensuelles :

Garanties	Saisonnalité	Tendance	Partie AR	Partie MA	Constante
RC Corporelle	$(1 - B^{12})$	NC *	$(1 - 0,57\ B^{3})$	NC	NC
RC matérielle	$(1 - B^{12})$	NC	$(1 - 0,55\ B^{3})$	NC	NC
Dommages	$(1 - B^{12})$	NC	NC	NC	NC
Bris de glace	$(1 - B^{12})$	NC	NC	NC	NC
Incendie	NC	NC	NC	NC	NC
Vol total	NC	$(1 - B)$	NC	$(1 - 0,87\ B)$	-0,0002
Vol partiel	$(1 - B^{12})$	$(1 - B)$	NC	$(1 - 0,82\ B)$	NC

* : Non concernée

Sur la garantie RC corporelle, on obtient un modèle de la même forme que dans le modèle complet, à savoir un AR(3). Les coefficients des deux modèles sont sensiblement différents : on a 0,51 dans le modèle complet et 0,57 dans le modèle tronqué.

Sur la garantie RC matérielle, on obtient aussi un AR(3) comme dans le modèle complet (coefficient égal à 0,51 dans le modèle complet et 0,55 dans le modèle tronqué).

Sur la garantie Dommages, on obtient un bruit blanc une fois la série désaisonnalisée, comme dans le modèle complet.

Pour la garantie Bris de glace, la série désaisonnalisée devient un bruit blanc, et non plus un ARIMA (1, 0,1) comme c'était le cas dans le modèle jusqu'à fin 2009. La structure du modèle n'est pas stable dans le temps : nous ne retiendrons pas ce modèle.

Pour la garantie Vol total, le modèle tronqué reste de la même forme que le modèle complet. Les coefficients ne sont cependant pas identiques : 0,80 dans le modèle complet contre 0,87 dans le modèle tronqué.

Sur la garantie Vol partiel, on retrouve le modèle $SARIMA_{12}[(0,1,1)(0,0,0)]$ avec des coefficients sensiblement identiques (0,79 pour le modèle complet et 0,82 pour le modèle tronqué).

Les prévisions obtenues pour l'année 2009 avec ces modèles tronqués sont représentées en annexe 8. On constate que les garanties dont les modèles tronqués prédisent assez bien l'année 2009 sont la RC matérielle, la RC corporelle et le Vol partiel.

Pour les séries de coûts de sinistres clos :

Avec les données tronquées à fin 2008, nous obtenons les modèles suivants :

Garanties	Saisonnalité	Tendance	Partie AR	Partie MA
RC Corporelle	$(1 - B^{12})$	NC *	$(1 - 0,52\ B^{3})$	NC
RC matérielle	$(1 - B^{12})$	$(1 - B)$	NC	NC
Dommages	$(1 - B^{12})$	$(1 - B)$	NC	NC
Bris de glace	NC	$(1 - B)$	NC	NC
Incendie	NC	NC	NC	NC
Vol total	NC	$(1 - B)$	NC	NC
Vol partiel	$(1 - B^{12})$	$(1 - B)$	NC	NC

* : Non concernée

En comparant les coefficients pour la série des coûts en RC corporelle, on constate que le modèle est robuste (les coefficients sont mêmes identiques). Et les données observées en 2009 restent dans l'intervalle de confiance de notre modèle.

Sur la RC matérielle, la série des coûts de sinistres clos qui nous donnait dans le modèle complet un modèle AR, devient ici un bruit blanc une fois retirées la tendance et la saisonnalité. Nous ne retiendrons pas ce modèle.

Pour les garanties restantes (à savoir le Dommages, le Bris de glace, le Vol total et le Vol partiel), il ne nous reste que des bruits blancs : la modélisation ARMA s'arrête alors dans ces cas précis.

Pour les séries de coûts de sinistres déclarés :

Avec les données tronquées à fin 2008, nous obtenons les modèles suivants :

Garanties	Saisonnalité	Tendance	Partie AR	Partie MA
RC matérielle	$(1 - B^{12})$	$(1 - B)$	$1 + 0{,}33\ B^2$	$1 - 0{,}70\ B^{12}$
Dommages	NC	NC	NC	NC
Bris de glace	$(1 - B^{12})$	$(1 - B)$	$1 - 0{,}77\ B^{12}$	NC
Incendie	$(1 - B^{12})$	NC	NC	NC
Vol total	$(1 - B^{12})$	$(1 - B)$	NC	$1 - 0,\ 73\ B$
Vol partiel	$(1 - B^{12})$	$(1 - B)$	$1 + 0{,}67\ B^{12}$	$1 - 0{,}55\ B$

* : Non concernée

Sur la RC matérielle, on retrouve un modèle de la même forme qu'avec les données complètes. Les coefficients sont quasiment les mêmes pour la partie AR et pour la partie MA.

Sur la garantie Dommages, nous n'avons pas testé de modèle car sur la série complète, on obtenait un bruit blanc après avoir retiré la tendance et la saisonnalité.

Sur la garantie Bris de glace, on obtient un modèle de la même forme que celui avec la série complète. Cependant le coefficient a varié de manière significative : ce modèle est instable.

Sur la garantie Incendie, la série désaisonnalisée devient un bruit blanc.

Sur la garantie Vol total, le modèle est identique en tout point à celui qu'on avait obtenu sur le modèle complet.

Sur la garantie Vol partiel, le modèle est stable mais les coefficients sont significativement différents à la fois pour la partie AR et la partie MA.

Les conclusions que l'on peut tirer sur l'ensemble des séries sont les suivantes : les modèles obtenus avec les données tronquées ne sont pas tous identiques aux modèles sur les données complètes. En effet, les coefficients des modèles SARIMA peuvent être significativement différents selon que les séries soient tronquées ou pas. Néanmoins les saisonnalités, tendances et ordres restent les mêmes pour quasiment toutes les séries.

Pour la RC corporelle, nous retenons cette famille de modèle avec les séries de fréquences et de coûts clos ; Pour la RC matérielle, avec les séries de fréquences et de coûts déclarés.

Pour les garanties Bris de glace, Dommages et Incendie, ce type de modèle n'est pas adapté.

Pour la garantie Vol partiel, on n'obtient un modèle qu'avec la série de fréquences. Pour la garantie Vol total, la modélisation fonctionne sur les séries de fréquences et de coûts déclarés.

V.2. Cas des modèles de régression avec erreurs ARMA

Pour les séries de fréquences mensuelles :

Les données tronquées nous donnent les résultats regroupés dans le tableau suivant :

Variables explicatives potentielles	RC corporelle	RC matérielle	Dommages	Bris de glace	Vol partiel	Vol total
Constante		-0,906	0,600		-0,030	0,052
Ensoleillement/100				0,040		
Parcours mensuel - autoroute interurbaine (21)				1,206		
Parcours mensuel - Autoroute et voie rapide urbaine (22)			-0,257			
Parcours mensuel - Route nationale interurbaine à caractéristiques autoroutière (23)				-1,240		-0,012
Vitesse de jour sur les RD		0,011				
Vitesse de jour sur les RN 2-2/100						0,018
Volume de Super	0,34				0,814	0,202
Volume de Gazole	0,06		1,157	2,623		
Volume de Super + Gazole		1,74	0,938			
AR (1)			0,266			
AR (3)	0,50	0,235			0,556	
AR (10)						-0,43
X1	- 0,023	-0,183	-0,095	-0,234		
X2	- 0,015	-0,111	-0,110		0,016	
X3	- 0,011	-0,112	-0,053		0,010	
X4	- 0,012	-0,123	-0,083		0,007	
X5	- 0,011	-0,074	-0,032		0,014	
X6		-0,030			0,011	
X7	- 0,008	-0,101	-0,074			
X8	- 0,009	-0,125	-0,140			
X9		-0,040	-0,015		0,009	
X10		-0,033			0,009	
X11					0,023	

Sur la garantie RC corporelle, on retrouve quasiment les mêmes coefficients que ceux du modèle complet. Ce sont les mêmes variables explicatives qui intègrent le modèle.

Sur la garantie RC matérielle, les variables que nous avions dans le modèle complet sont toujours significatives, avec quasiment les mêmes coefficients et nous donnent un bon ajustement avec les données observées sur l'année 2009. On tire les mêmes conclusions pour les garanties Dommages, Bris de glace, Vol partiel et Vol total.

Les prévisions obtenues pour l'année 2009, avec nos données tronquées à fin 2008 sont disponibles en annexe 9.

V.3. Cas des modèles UCM

Pour ces modèles, nous avons testé la stabilité de la structure de nos séries, en termes de composantes intégrant le modèle. Nous avons également comparé le réel 2009 avec les prévisions 2009 obtenues avec les données tronquées. Les résultats obtenus sont détaillés dans cette partie, avec les graphiques des prévisions obtenues.

Modèles saturés :

Séries de fréquences mensuelles

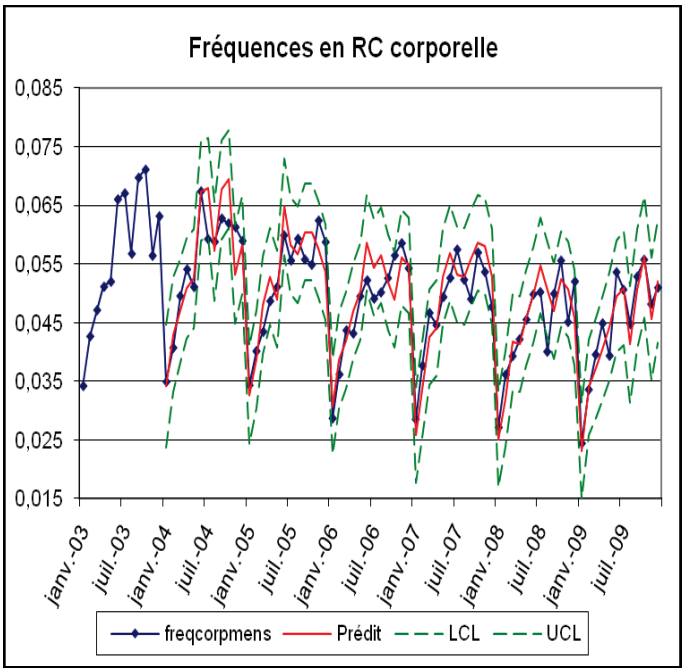
Pour la garantie RC corporelle :

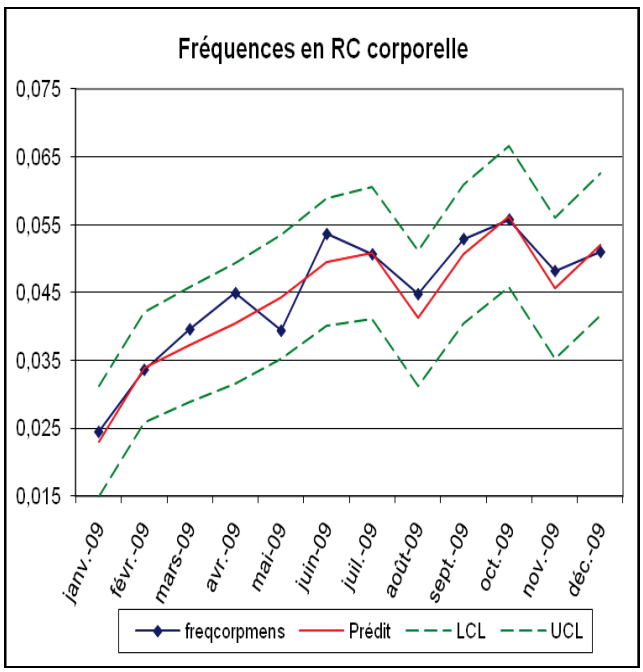
Avec la série tronquée à fin 2008, nous obtenons le modèle suivant :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	998.78	<.0001
Season	11	467.39	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00001767
Root Mean Squared Error	0.00420
Mean Absolute Percentage Error	7.12748
Maximum Percent Error	13.37399
R-Square	0.78469
Adjusted R-Square	0.78098
Random Walk R-Square	0.81216
Amemiya's Adjusted R-Square	0.76984
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 60	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	0.00000144	7.02979E-7	2.05	0.0402
Season	**Error Variance**	0.00000569	1.53362E-6	3.71	0.0002





On constate que le modèle est stable par rapport au modèle complet (mêmes composantes significatives). Aussi, les prévisions pour l'année 2009 s'ajustent plutôt bien au réel 2009 : le graphique de droite nous le montre clairement.

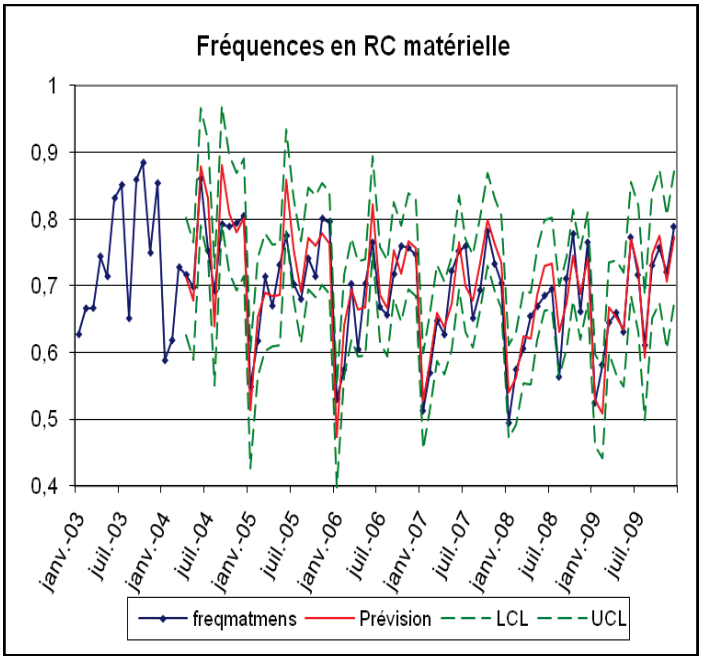
Pour la garantie RC matérielle :

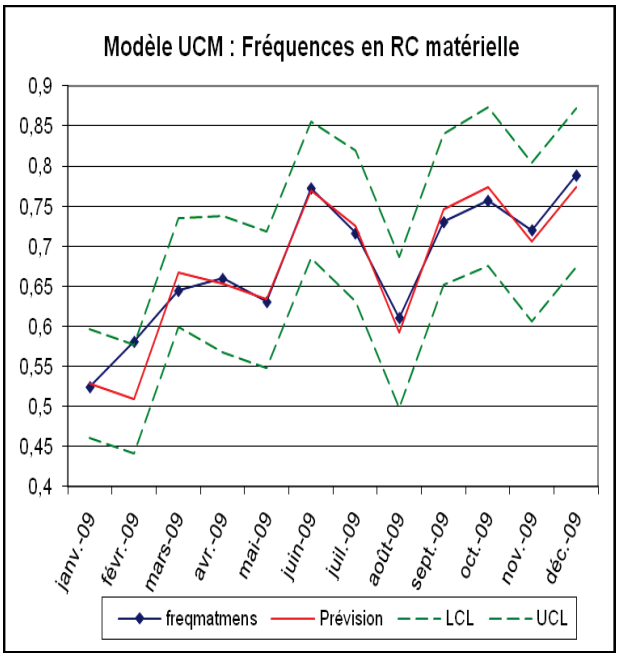
Avec la série tronquée en fin 2008, nous avons obtenu le modèle suivant :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	2.12	0.1459
Level	1	1658.77	<.0001
Season	11	677.16	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00148
Root Mean Squared Error	0.03841
Mean Absolute Percentage Error	4.76315
Maximum Percent Error	10.60505
R-Square	0.77281
Adjusted R-Square	0.76868
Random Walk R-Square	0.91662
Amemiya's Adjusted R-Square	0.75629
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 57	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00101	0.0001891	5.34	<.0001
DepLag	**Phi_1**	0.77264	0.08237	9.38	<.0001





La saisonnalité est déterministe (variance fixée à 0). Le retard d'ordre 3 est significatif tout comme dans le modèle complet. Notre modèle est stable : on retrouve les mêmes composantes. Le zoom sur la partie 2009 (réel vs prédit) montre l'adéquation du modèle avec la réalité sur l'année 2009.

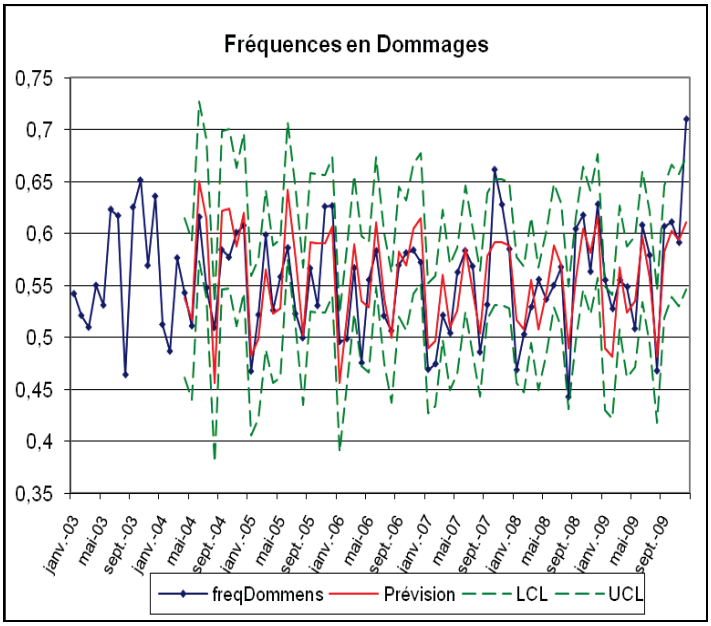
Pour la garantie Dommages :

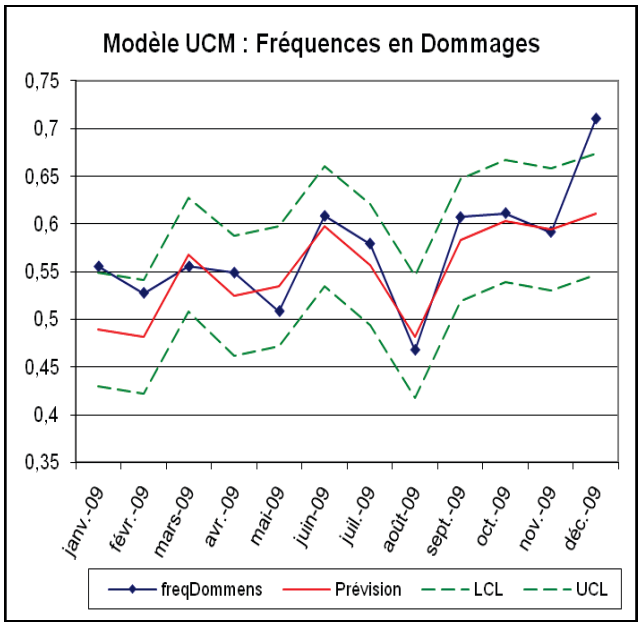
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	0.69	0.4059
Level	1	9930.03	<.0001
Season	11	211.38	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00110
Root Mean Squared Error	0.03315
Mean Absolute Percentage Error	5.06994
Maximum Percent Error	10.58300
R-Square	0.54569
Adjusted R-Square	0.53743
Random Walk R-Square	0.88091
Amemiya's Adjusted R-Square	0.51265
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 57	

On a introduit un retard d'ordre 3 dans le modèle, comme dans le modèle complet. De même, la tendance et la saisonnalité sont déterministes. Le modèle est stable. Mais la forte hausse de décembre 2009 n'avait pas été prédite par notre modèle, qui prévoyait certes une hausse mais pas d'une telle ampleur.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00076461	0.0001432	5.34	<.0001
DepLag	**Phi_1**	0.39714	0.11693	3.40	0.0007



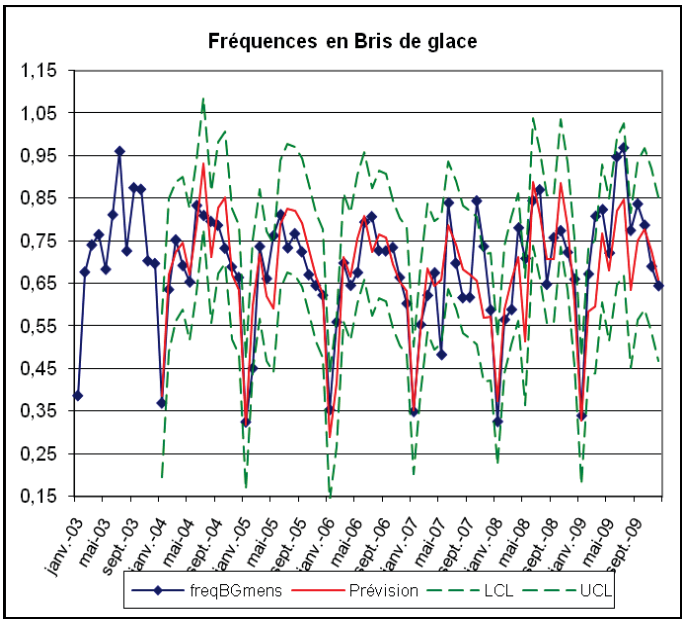


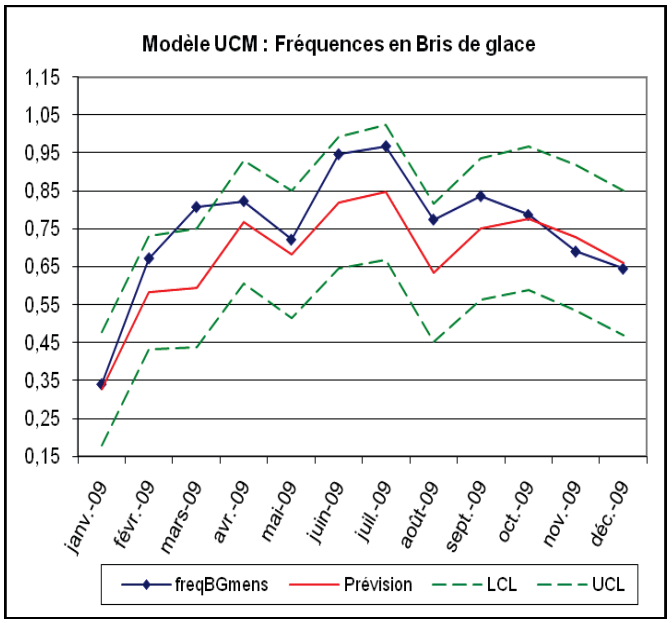
Pour la garantie Bris de glace :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	560.57	<.0001
Season	11	277.93	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00611
Root Mean Squared Error	0.07815
Mean Absolute Percentage Error	9.13543
Maximum Percent Error	25.63320
R-Square	0.68362
Adjusted R-Square	0.67910
Random Walk R-Square	0.73477
Amemiya's Adjusted R-Square	0.66554
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 72	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	0.00085259	0.0003629	2.35	0.0188
Season	**Error Variance**	0.00151	0.0004412	3.43	0.0006





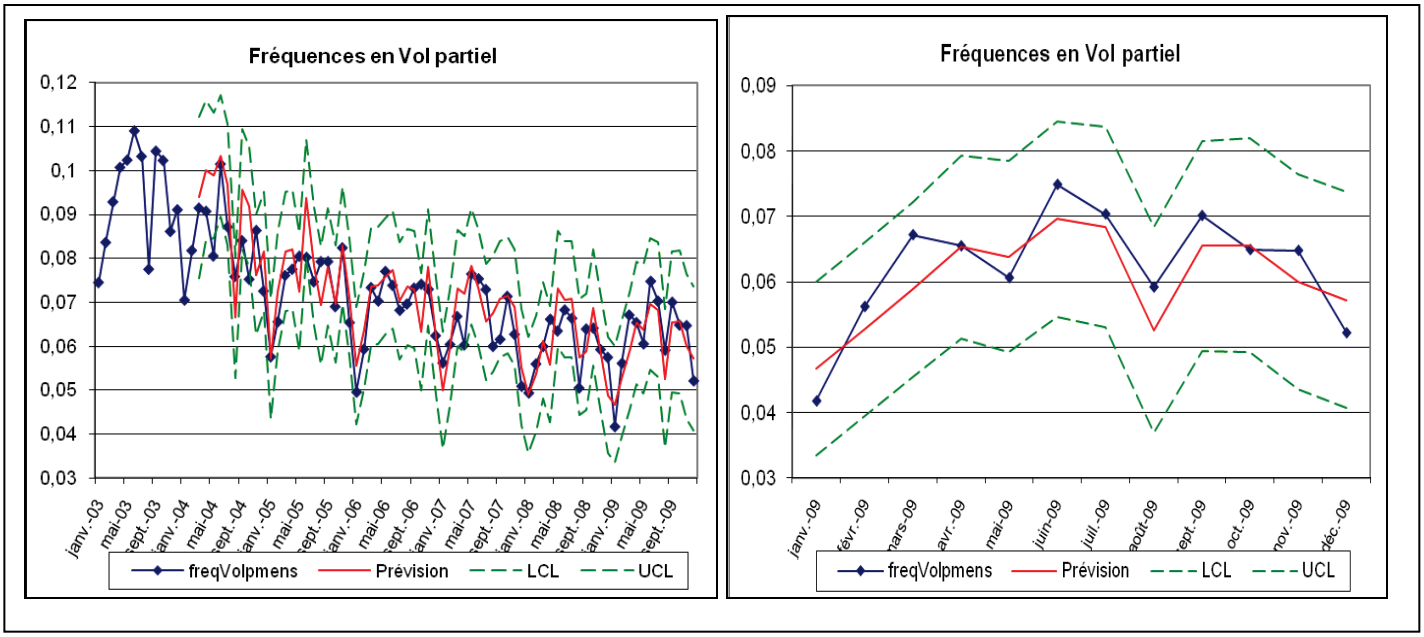
Les composantes dans le modèle tronqué sont les mêmes que celles qu'on avait dans le modèle complet. On constate une forte hausse en 2009, que notre modèle n'avait pas prédite. Néanmoins, les vraies valeurs de 2009 restent dans l'intervalle de confiance de notre modèle. L'ajustement du modèle tronqué aux valeurs réelles est satisfaisant.

Pour la garantie Vol partiel :

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	0.00000513	2.02543E-6	2.53	0.0113
Season	**Error Variance**	0.00001305	3.41218E-6	3.82	0.0001
DepLag	**Phi_1**	-0.25425	0.13752	-1.85	0.0645

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	986.18	<.0001
Season	11	143.28	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00004815
Root Mean Squared Error	0.00694
Mean Absolute Percentage Error	7.84583
Maximum Percent Error	15.54085
R-Square	0.59724
Adjusted R-Square	0.58259
Random Walk R-Square	0.78390
Amemiya's Adjusted R-Square	0.55330
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 58	



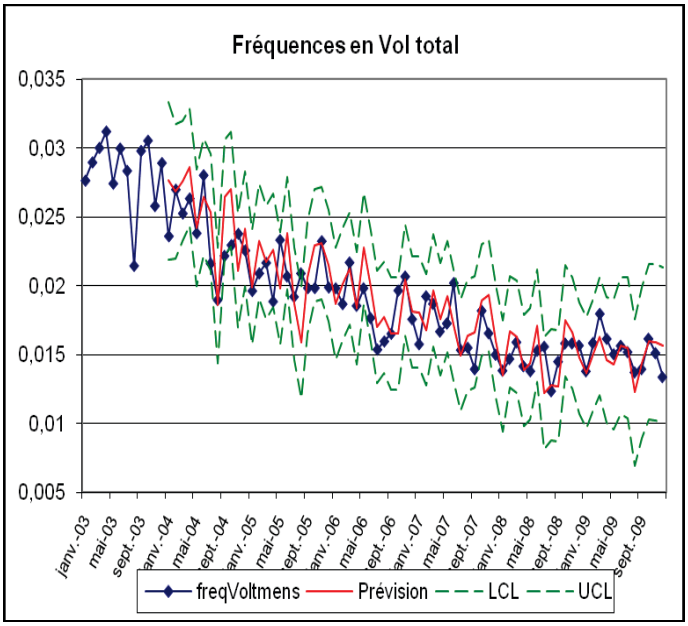
Les prévisions du modèle tronqué et l'observé sur l'année 2009 sont satisfaisantes. Néanmoins, le modèle tronqué intègre un retard qui n'était pas présent dans le modèle complet. La structure du modèle a changé. Nous ne retiendrons donc pas ce modèle pour les fréquences en vol partiel.

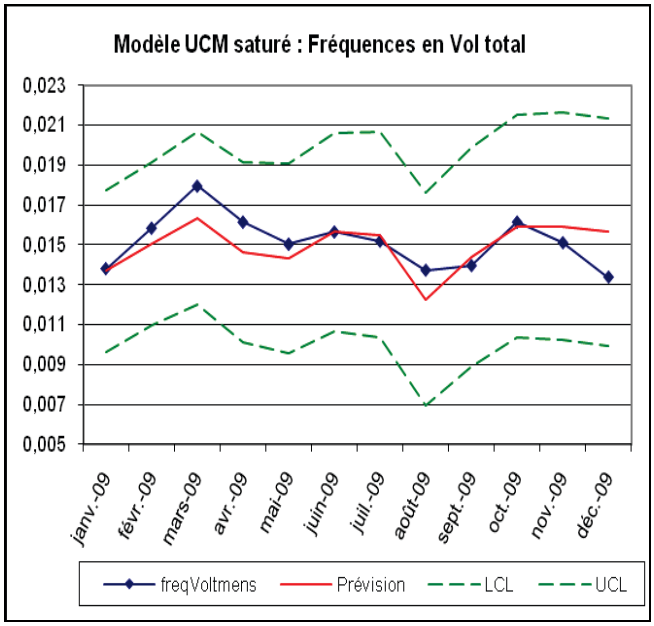
Pour la garantie Vol total :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	399.00	<.0001
Season	11	54.08	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00000452
Root Mean Squared Error	0.00213
Mean Absolute Percentage Error	8.91772
Maximum Percent Error	23.97199
R-Square	0.64725
Adjusted R-Square	0.64117
Random Walk R-Square	0.69279
Amemiya's Adjusted R-Square	0.62292
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 60	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	5.064984E-7	1.79534E-7	2.82	0.0048
Season	**Error Variance**	0.00000119	2.92832E-7	4.05	<.0001





Les prévisions obtenues avec notre modèle et le réel 2009 s'ajustent come on peut le voir sur le graphique de droite. Le modèle avec les données tronquées intègrent les mêmes composantes significatives que le modèle complet. On valide cette modélisation.

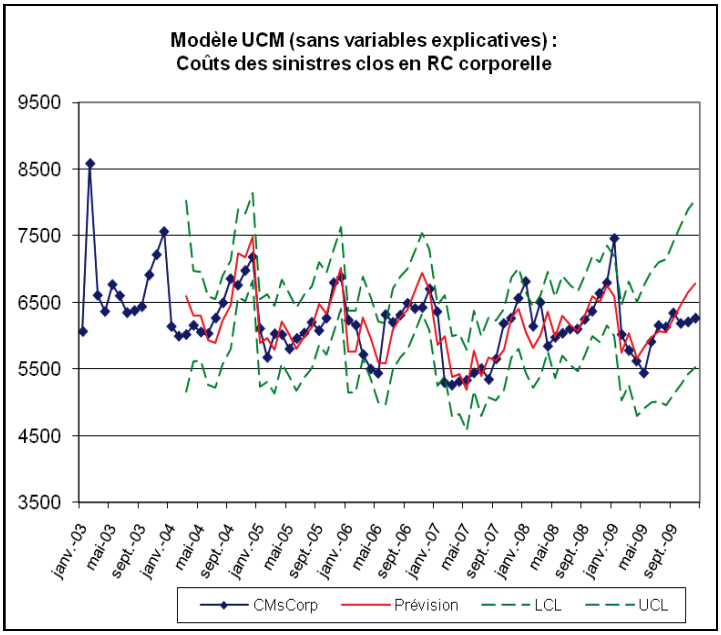
Séries de coûts de sinistres clos

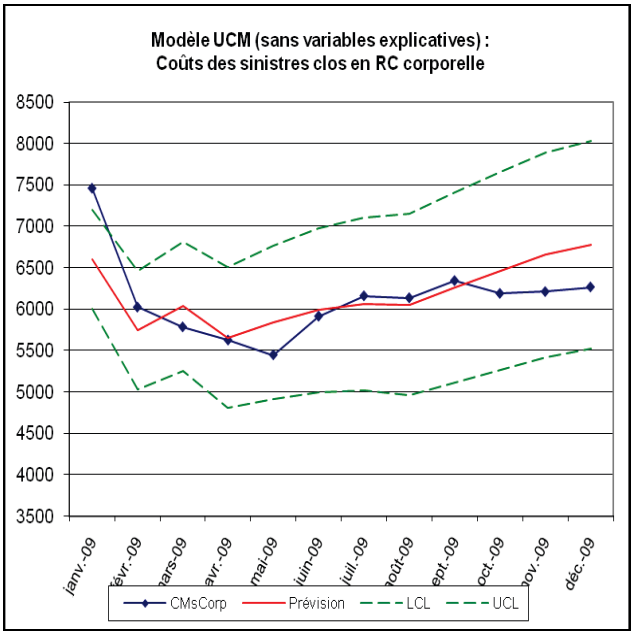
Pour la garantie RC corporelle :

Le modèle saturé sur les données tronquées à fin 2008 nous donne les résultats ci-dessous :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	2940.61	<.0001
Season	11	76.87	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	103236
Root Mean Squared Error	321.30361
Mean Absolute Percentage Error	4.18563
Maximum Percent Error	11.61671
R-Square	0.50883
Adjusted R-Square	0.49097
Random Walk R-Square	0.85834
Amemiya's Adjusted R-Square	0.45524
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 58	





Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	42841	12328.7	3.47	0.0005
Season	**Error Variance**	6942.72566	3679.1	1.89	0.0591
DepLag	**Phi_1**	-0.17052	0.08383	-2.03	0.0419

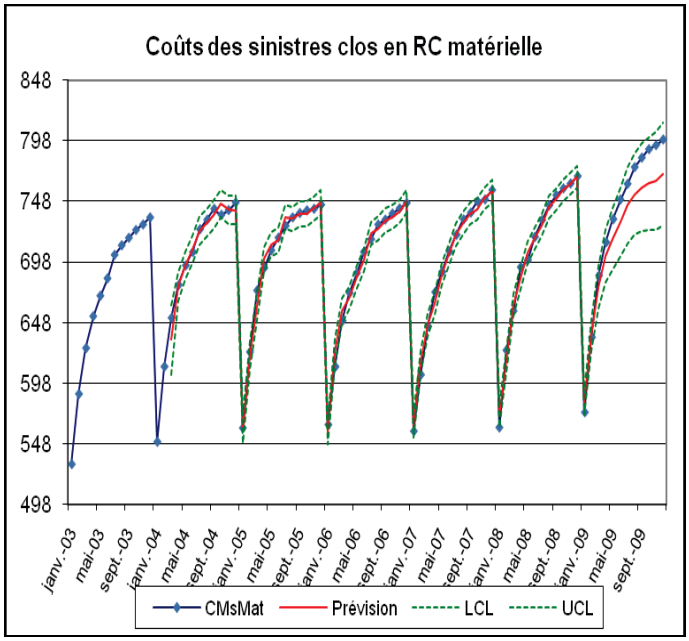
On a introduit le retard d'ordre 2 dans le modèle tronqué, tout comme dans le modèle complet. Les composantes significatives sont restées les mêmes. Le modèle prédit pour l'année 2009 s'ajuste assez bien avec le réel observé sur 2009.

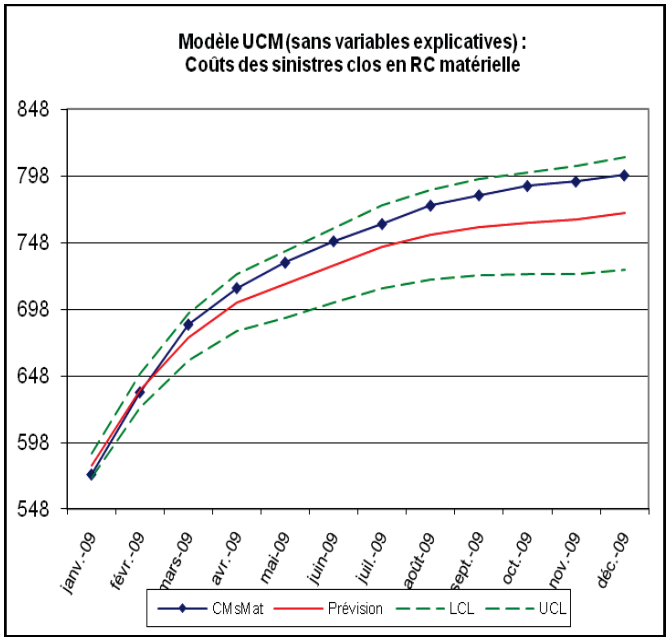
Pour la garantie RC matérielle :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	61068.7	<.0001
Season	11	14089.4	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	30.33944
Root Mean Squared Error	5.50812
Mean Absolute Percentage Error	0.59917
Maximum Percent Error	2.75118
R-Square	0.98981
Adjusted R-Square	0.98963
Random Walk R-Square	0.99698
Amemiya's Adjusted R-Square	0.98908
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 58	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	18.10387	3.36180	5.39	<.0001
DepLag	**Phi_1**	0.37893	0.12139	3.12	0.0018





L'ajustement est satisfaisant. Les composantes du modèle tronqué sont les mêmes que celles du modèle complet.

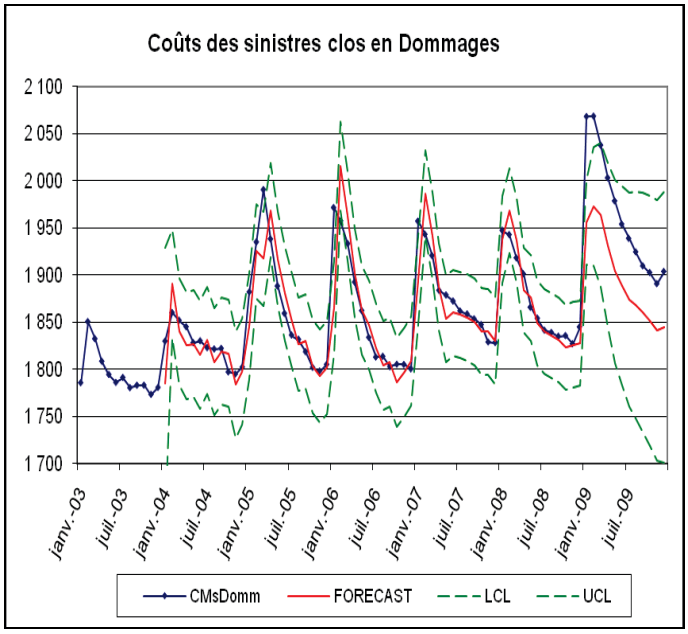
Pour la garantie Dommages :

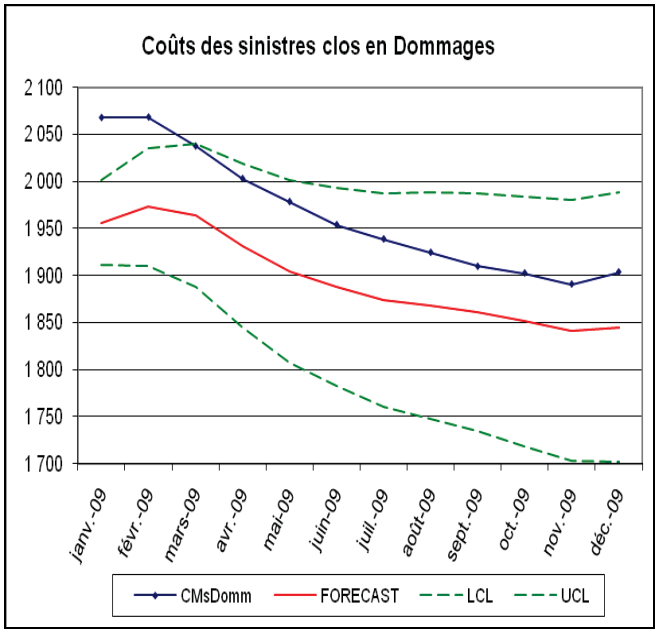
Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Level	1	46769.2	<.0001
Season	11	177.48	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	654.26033
Root Mean Squared Error	25.57851
Mean Absolute Percentage Error	0.89876
Maximum Percent Error	5.33147
R-Square	0.74552
Adjusted R-Square	0.74552
Random Walk R-Square	0.98842
Amemiya's Adjusted R-Square	0.73689
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 60	

Le modèle obtenu intègre les mêmes composantes significatives que le modèle complet.

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	446.01415	81.43067	5.48	<.0001





L'ajustement de notre modèle avec le réel 2009 est satisfaisant, même si l'intervalle de confiance s'élargit beaucoup vers les derniers mois de l'année.

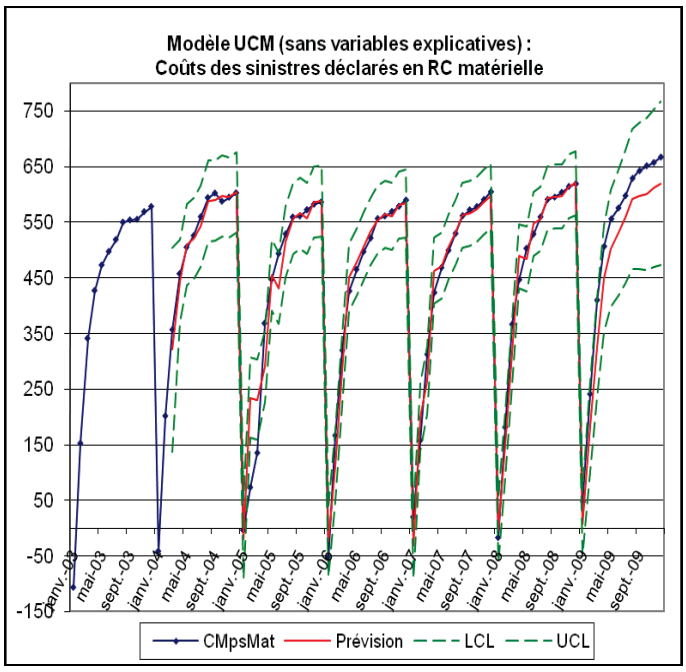
Séries de coûts de sinistres déclarés

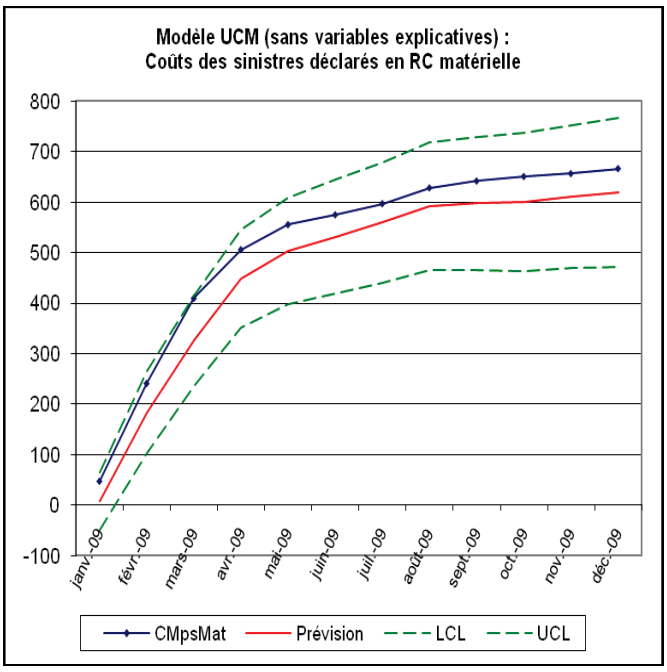
Pour la garantie RC matérielle :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	2838.16	<.0001
Season	11	3164.08	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	1174.74445
Root Mean Squared Error	34.27455
Mean Absolute Percentage Error	24.74193
Maximum Percent Error	243.69756
R-Square	0.96458
Adjusted R-Square	0.96395
Random Walk R-Square	0.96464
Amemiya's Adjusted R-Square	0.96205
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 58	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	731.84461	135.90013	5.39	<.0001
DepLag	**Phi_1**	-0.31219	0.11462	-2.72	0.0065





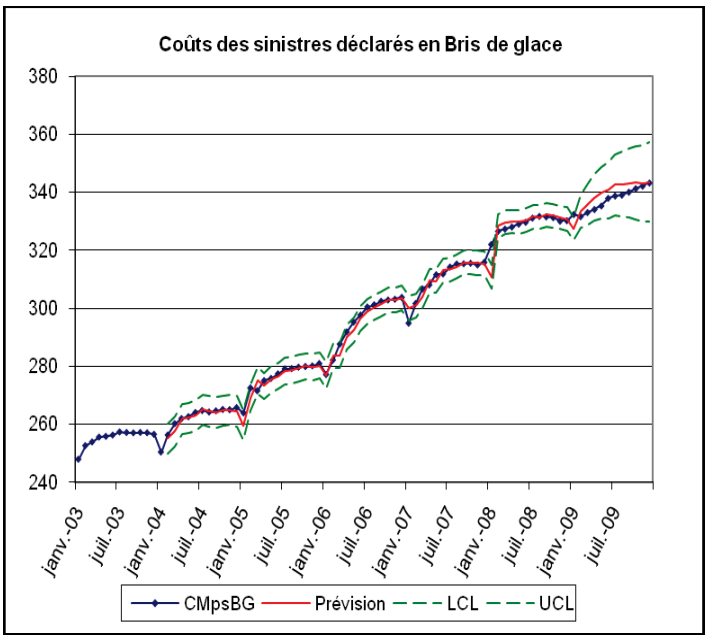
Le modèle obtenu intègre les mêmes composantes significatives que le modèle complet, avec le retard d'ordre 2. L'ajustement de nos prévisions au réel 2009 est satisfaisant.

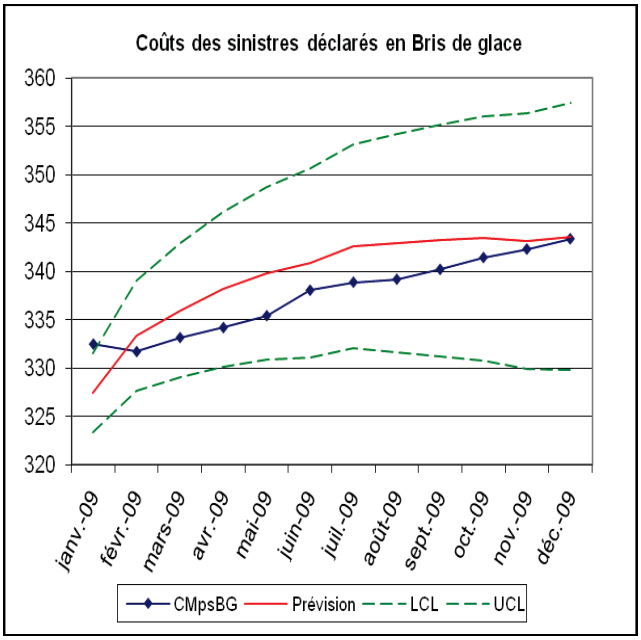
Pour la garantie Bris de glace :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	180256	<.0001
Slope	1	24.33	<.0001
Season	11	76.08	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	4.76935
Root Mean Squared Error	2.18389
Mean Absolute Percentage Error	0.45459
Maximum Percent Error	3.51676
R-Square	0.99174
Adjusted R-Square	0.99174
Random Walk R-Square	0.99563
Amemiya's Adjusted R-Square	0.99146
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	3.55248	0.65406	5.43	<.0001





Les composantes significatives dans le modèle complet le sont aussi dans le modèle tronqué.

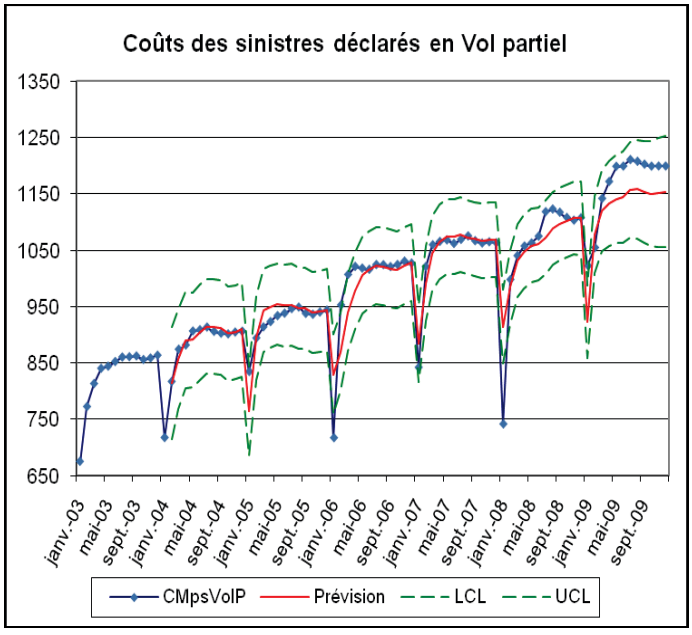
Le modèle prédit de manière satisfaisante les valeurs sur l'année 2009.

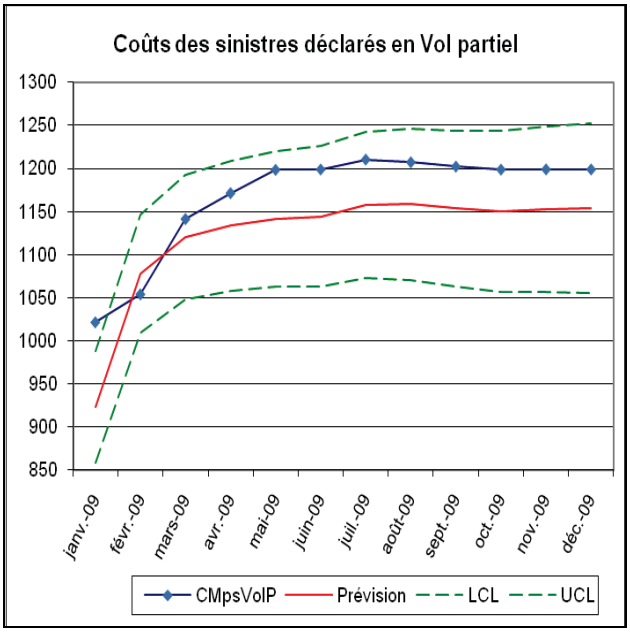
Pour la garantie Vol partiel :

Analyse de significativité des composantes (basée sur l'état final)			
Composante	**DDL**	**Khi-2**	**Pr > Khi-2**
Irregular	1	0.00	0.9871
Level	1	5104.07	<.0001
Slope	1	9.17	0.0025
Season	11	283.88	<.0001

Statistiques d'ajustement basées sur les résidus	
Mean Squared Error	1214.20180
Root Mean Squared Error	34.84540
Mean Absolute Percentage Error	2.03059
Maximum Percent Error	8.60927
R-Square	0.85965
Adjusted R-Square	0.85719
Random Walk R-Square	0.80773
Amemiya's Adjusted R-Square	0.84981
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Valeur estimée**	**Erreur type approchée**	**Valeur du test t**	**Approx. de Pr > |t|**
Irregular	**Error Variance**	592.16521	160.50184	3.69	0.0002
Level	**Error Variance**	112.51027	88.04509	1.28	0.2013





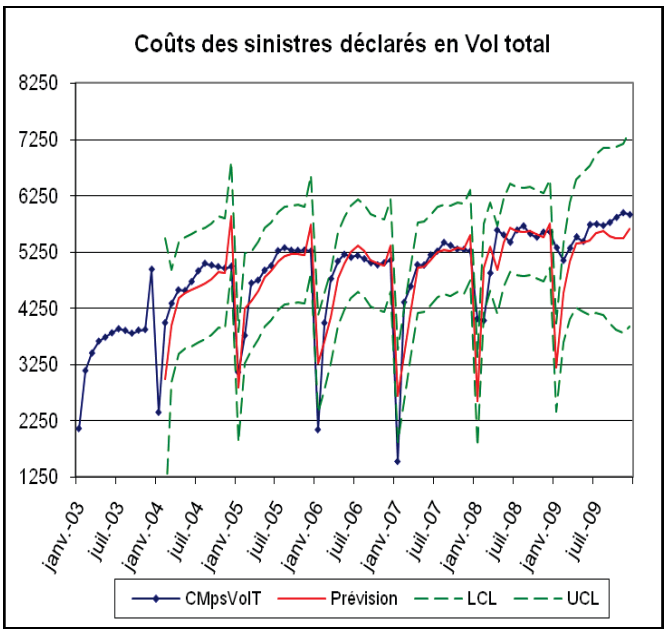
Les composantes significatives dans le modèle complet n'intègre pas le cycle de 31 mois que nous avions dans le modèle complet. Néanmoins, les prévisions pour l'année 2009 englobent les valeurs réellement observées cette année-là.

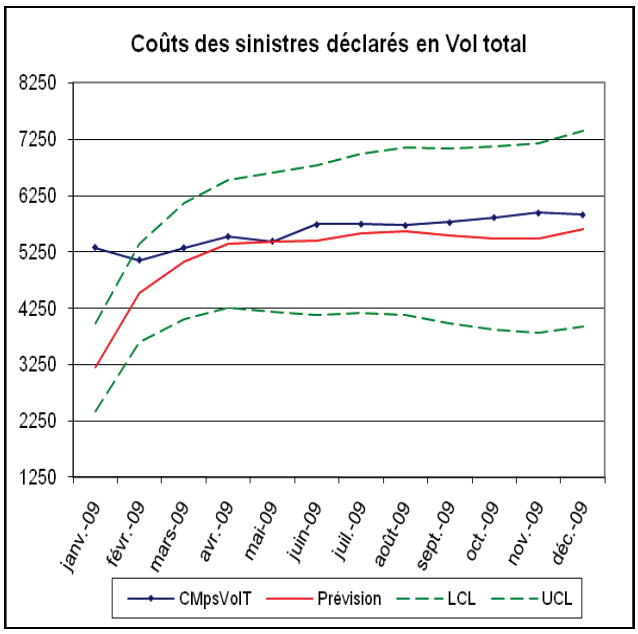
Pour la garantie Vol total :

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	2525.84	<.0001
Season	11	299.44	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	197316
Root Mean Squared Error	444.20305
Mean Absolute Percentage Error	7.75770
Maximum Percent Error	36.07602
R-Square	0.66573
Adjusted R-Square	0.65987
Random Walk R-Square	0.80325
Amemiya's Adjusted R-Square	0.64228
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	135352	24920.3	5.43	<.0001
DepLag	**Phi_1**	-0.49374	0.11278	-4.38	<.0001





Le modèle tronqué conserve les mêmes composantes que le modèle complet. Les prévisions pour l'année 2009 collent aux observations de cette année là.

Modèles non saturés :

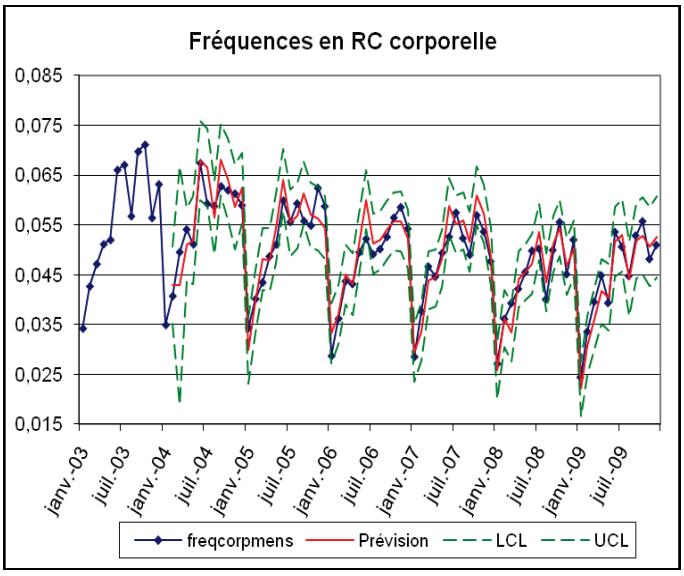
Séries de fréquences mensuelles

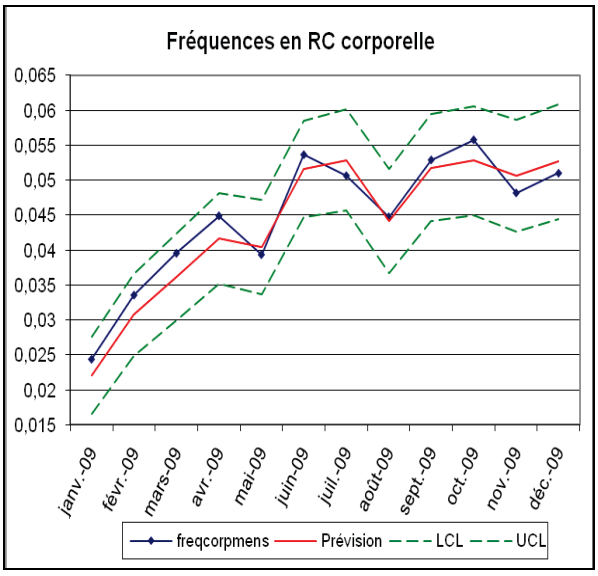
Pour la garantie RC corporelle :

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00000441	1.1394E-6	3.87	0.0001
Level	**Error Variance**	8.994442E-7	5.71558E-7	1.57	0.1156
Super_gazole_m3	**Coefficient**	0.16361	0.02055	7.96	<.0001

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	0.43	0.5136
Level	1	8.20	0.0042
Season	11	494.72	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00001093
Root Mean Squared Error	0.00331
Mean Absolute Percentage Error	5.48126
Maximum Percent Error	14.88722
R-Square	0.86306
Adjusted R-Square	0.86066
Random Walk R-Square	0.89341
Amemiya's Adjusted R-Square	0.85345
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	





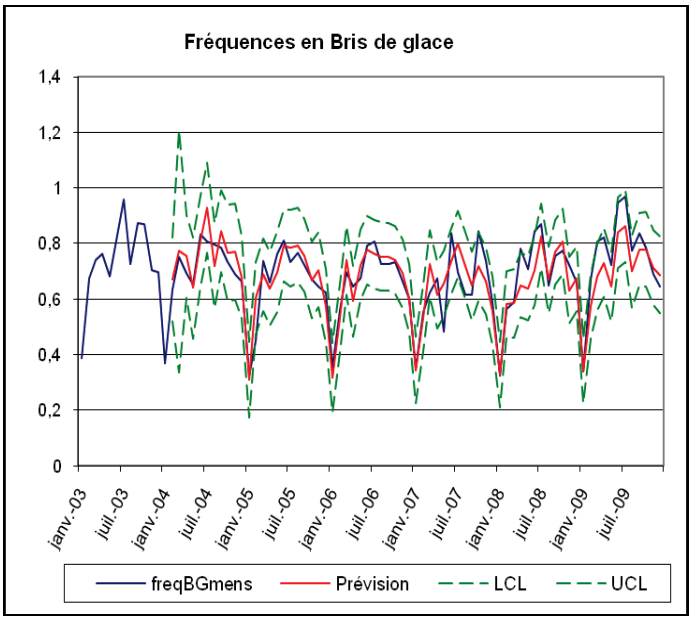
Les composantes significatives sont les mêmes que dans le modèle complet. L'ajustement par rapport au réel observé en 2009 est satisfaisant.

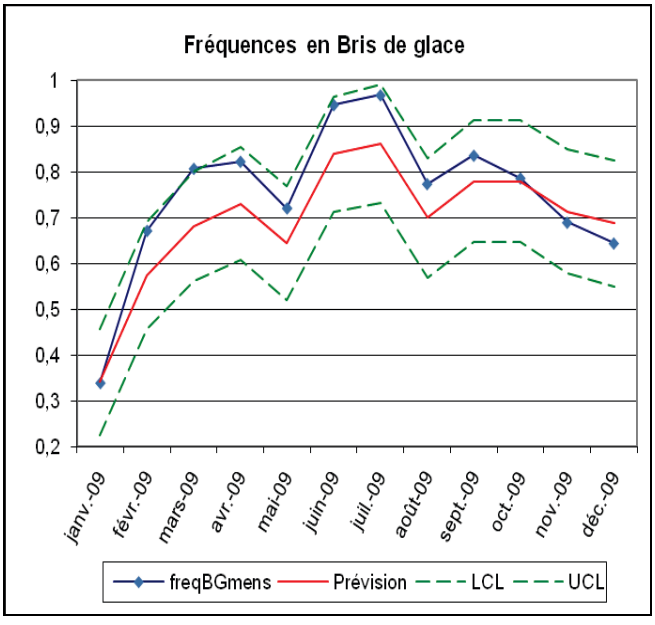
Pour la garantie Bris de glace :

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	0.00232	0.0004897	4.73	<.0001
Level	**Error Variance**	0.00014687	0.0001042	1.41	0.1588
Super_gazole_m3	**Coefficient**	1.95342	0.45335	4.31	<.0001

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	0.10	0.7479
Level	1	0.45	0.5030
Season	11	194.85	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	0.00417
Root Mean Squared Error	0.06454
Mean Absolute Percentage Error	7.48650
Maximum Percent Error	16.60261
R-Square	0.73816
Adjusted R-Square	0.73357
Random Walk R-Square	0.83669
Amemiya's Adjusted R-Square	0.71979
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	





Les prévisions sur l'année 2009 sont en-dessous des valeurs réellement observées. Néanmoins, on reste dans la bande de confiance du modèle. Ce modèle est validé.

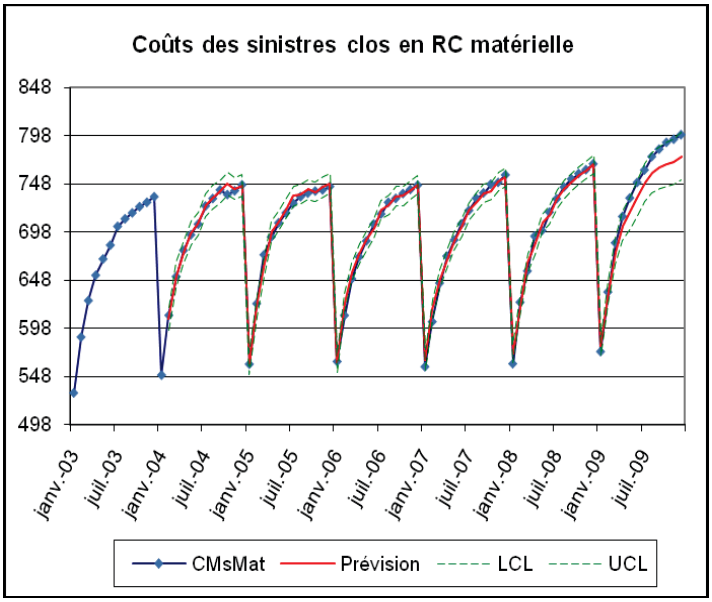
Séries de coûts de sinistres clos

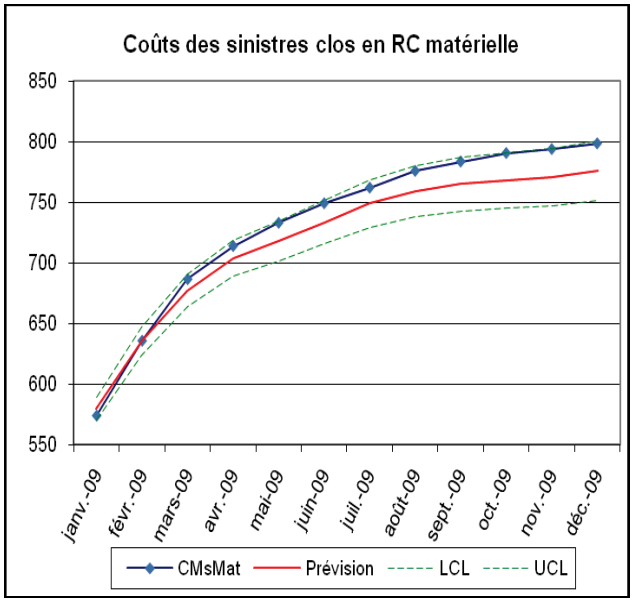
Pour la garantie RC matérielle :

Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Irregular	**Error Variance**	3.95267	2.06788	1.91	0.0559
Level	**Error Variance**	11.23689	3.67596	3.06	0.0022
ipc_piece_access	**Coefficient**	2.16374	1.01963	2.12	0.0338

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	0.00	0.9772
Level	1	11.49	0.0007
Season	11	8997.44	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	25.34220
Root Mean Squared Error	5.03410
Mean Absolute Percentage Error	0.55219
Maximum Percent Error	1.84099
R-Square	0.99172
Adjusted R-Square	0.99157
Random Walk R-Square	0.99720
Amemiya's Adjusted R-Square	0.99114
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	





Les prévisions qu'on obtient avec notre modèle tronqué sont en-dessous du réel observé sur l'année 2009 ; mais on reste dans la bande de confiance du modèle. Ce modèle est validé.

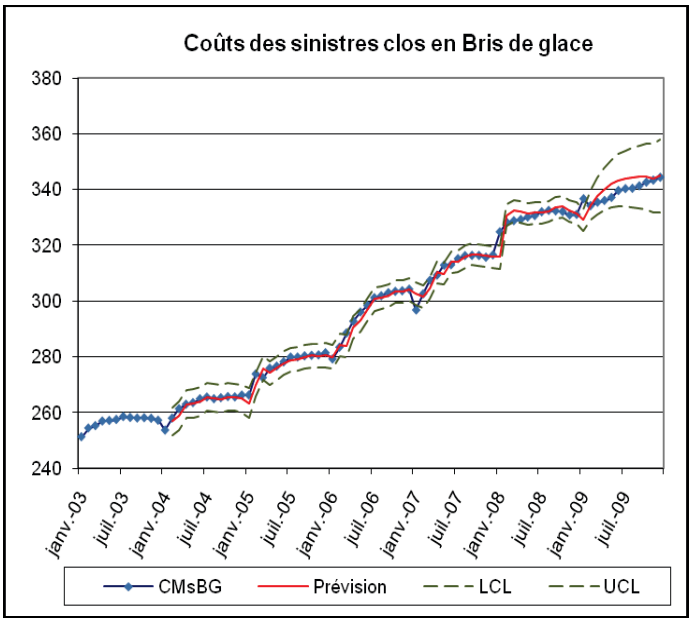
Pour la garantie Bris de glace :

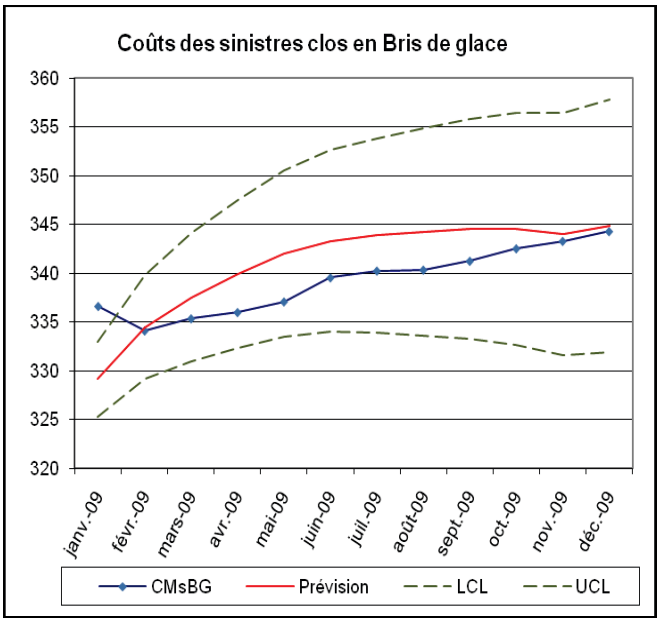
Final Estimates of the Free Parameters					
Composante	**Paramètre**	**Estimation**	**Erreur std approchée**	**Valeur du test t**	**Pr. Approx. > |t|**
Level	**Error Variance**	3.19327	0.58793	5.43	<.0001
ipc_repar_veh_perso	**Coefficient**	2.45247	0.40906	6.00	<.0001

Analyse de signification des composantes (basée sur l'état final)			
Composante	**DF**	**Khi 2**	**Pr > Khi 2**
Irregular	1	.	.
Level	1	0.30	0.5857
Season	11	70.14	<.0001

Statistiques d'ajustement basé sur les résidus	
Mean Squared Error	4.47431
Root Mean Squared Error	2.11526
Mean Absolute Percentage Error	0.46672
Maximum Percent Error	2.76569
R-Square	0.99223
Adjusted R-Square	0.99223
Random Walk R-Square	0.99595
Amemiya's Adjusted R-Square	0.99196
Nombre de résidus non manquants utilisés pour calculer les stat. d'ajustement = 59	

Le modèle sur les données tronquées conserve les mêmes composantes significatives que celui sur les données complètes. L'ajustement de nos prévisions sur les valeurs observées en 2009 est satisfaisant. Ce modèle est validé.





V.4. Cas des modèles de régression linéaire multiple

Avec la MM-estimation :

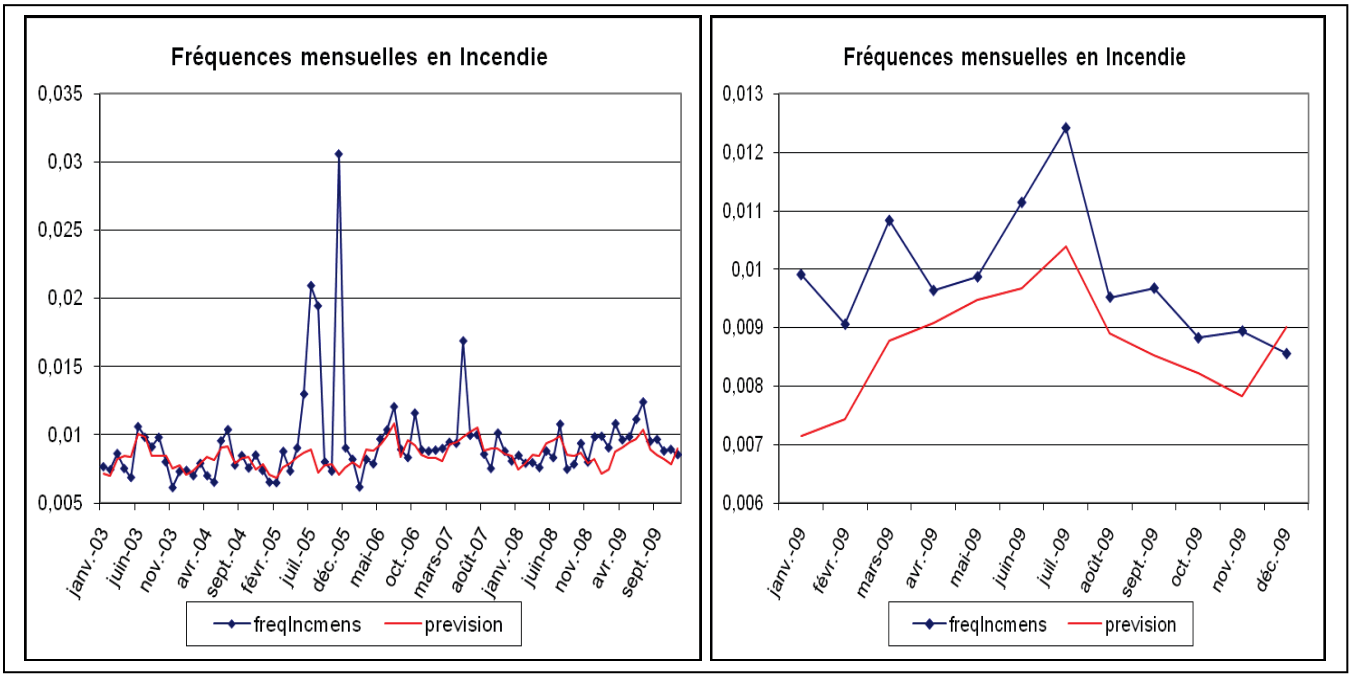
Avec les données tronquées à fin 2008, nous avons refait une régression multiple et avec ce modèle, nous avons effectué des prévisions sur l'année 2009.

Le modèle obtenu, avec un R carré de 0,81 est :

Incendie = 0,0060×Parcours mensuel sur les autoroutes et voies rapides urbaines - 0,0079×Parcours mensuel sur route nationale interurbaine à caractéristiques autoroutières – 0,0089×Parcours mensuel sur les autres routes nationales + 0,0019×Parcours mensuel total sur autoroutes + 0,0001×Températures + ε_t.

Ce modèle est stable : il intègre les mêmes variables explicatives que le modèle complet. Les coefficients des modèles tronqué et complet ne sont pas très différents.

Les résultats obtenus sont illustrés par les graphiques ci-dessous :



On constate que les prévisions obtenues avec le modèle tronqué sont en dessous des fréquences effectivement obtenues sur 2009. Avec une erreur moyenne (MAPE) de 11,6%, l'erreur que nous faisons est assez faible pour des fréquences de l'ordre de 0,009.

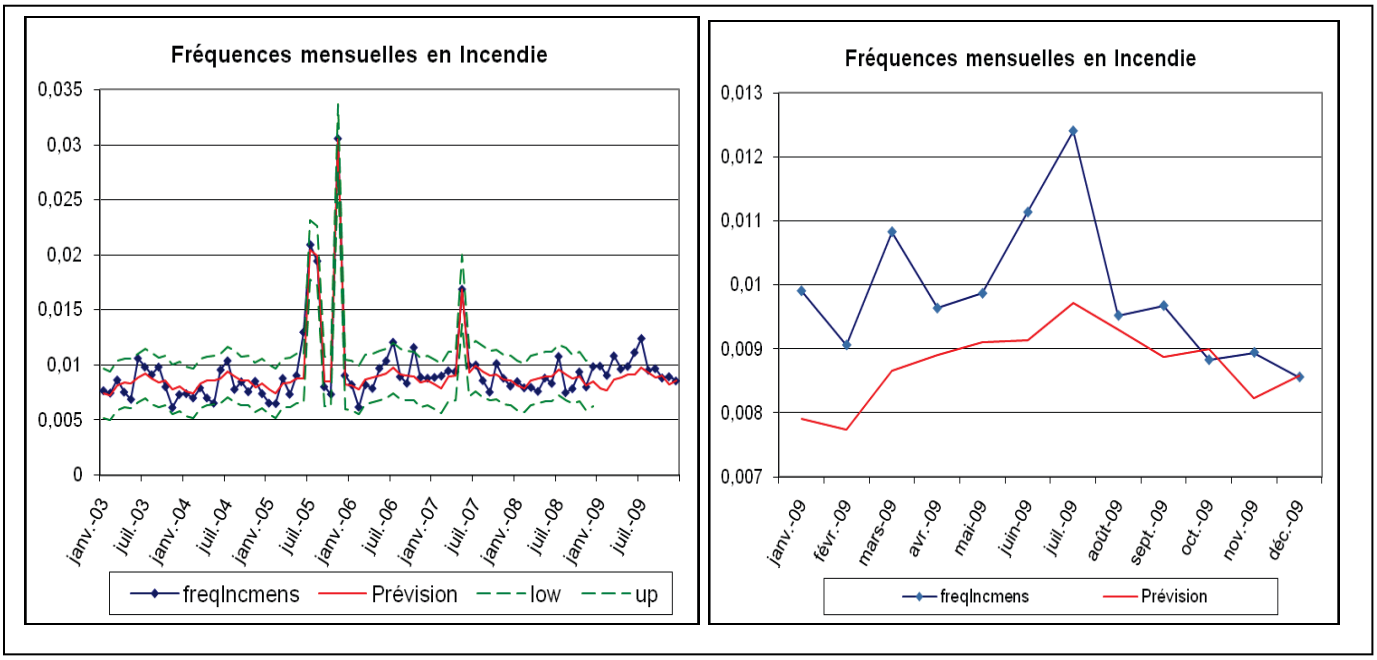
Avec l'introduction des variables d'intervention :

Avec les observations tronquées, le modèle obtenu est :

Incendie = 0,00182 Parcours mensuel sur autoroutes et voies rapides urbaines + 0,00028 Parcours moyen mensuel sur le total des autoroutes + 0,0108 pulse1 + 0,0224 pulse2 + 0,0077 pulse3.

Nous avons un R^2 de 0,98.

Les variables qui interviennent dans le modèle complet sont les mêmes que dans le modèle tronqué, avec des coefficients pas très différents.



Nous n'avons pas obtenu d'intervalle de confiance sur les prévisions 2009 de ce modèle. L'erreur moyenne (MAPE) que nous commettons est de 9,9%. Ce qui pour des fréquences de l'ordre de 0,009 est plutôt faible.

VI. PREVISIONS ET CHOIX DU MEILLEUR MODELE

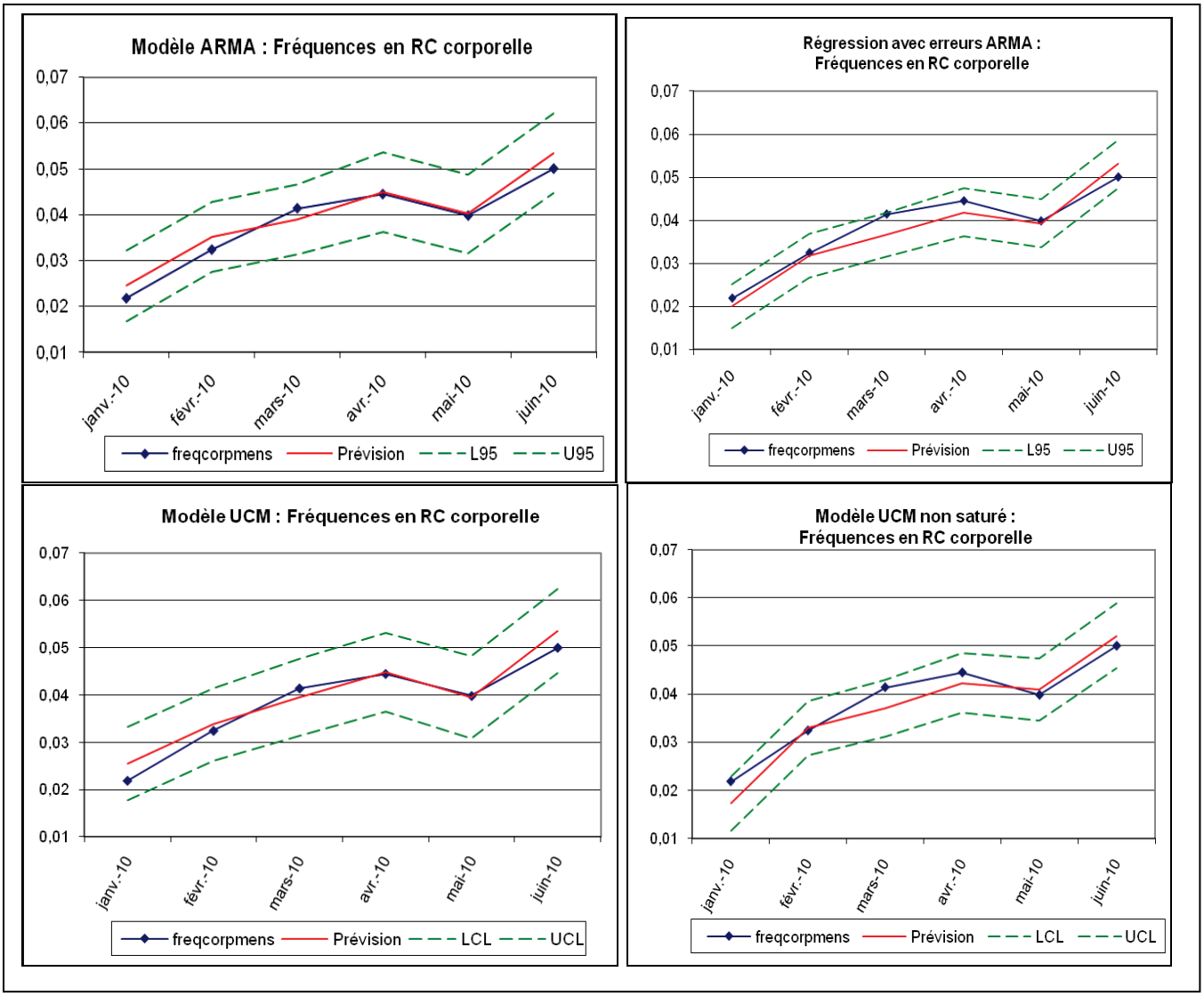
La validation de nos modèles terminée, nous devons décider du modèle que nous allons conserver pour chaque série étudiée. Pour choisir, nous allons comparer les modèles avec des critères prédictifs tels que le RMSE (Root Mean Square Error), le MAPE (Mean Absolute Percentage Error) et le MPE (Maximum Percentage Error). La méthode de calcul de ces indicateurs est disponible en annexe 12.
Ensuite, nous effectuerons des prévisions sur le premier semestre 2010, prévisions que nous comparerons au réel obtenu sur le premier semestre 2010.

VI.1. La garantie RC corporelle

Pour la série de fréquences mensuelles :
Les modèles qui s'ajustaient le mieux à nos séries étaient : le modèle $SARIMA_{12}[(3,0,0)(0,0,0)]$, le modèle de régression avec erreurs ARMA et le modèle UCM saturé et le modèle UCM non saturé.
La comparaison de ces différents modèles est résumée dans le tableau ci-dessous :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(3) $SARIMA_{12}[(3,0,0)(0,0,0)]$	0,00390	6,5%	0,010	25,69%
(1) Régression avec erreurs ARMA	0,00245	3,4%	0,007	13,42%
(4) UCM saturé	0,00397	6,7%	0,010	13,3%
(2) UCM non saturé	0,00316	5,4%	0,008	15,8%

Au vu de ce tableau, le meilleur modèle serait le modèle de régression avec erreurs ARMA. Vu que nous disposions des données sur le premier semestre 2010, nous avons réalisé des prévisions sur l'année 2010 avec chacun des modèles, pour voir celui qui se rapprochait le plus de la réalité. Les graphiques ci-dessous illustrent les résultats obtenus :



On remarque que les intervalles de confiance du modèle de régression avec erreurs ARMA sont les moins larges. Au final, c'est le modèle que nous avons retenu.

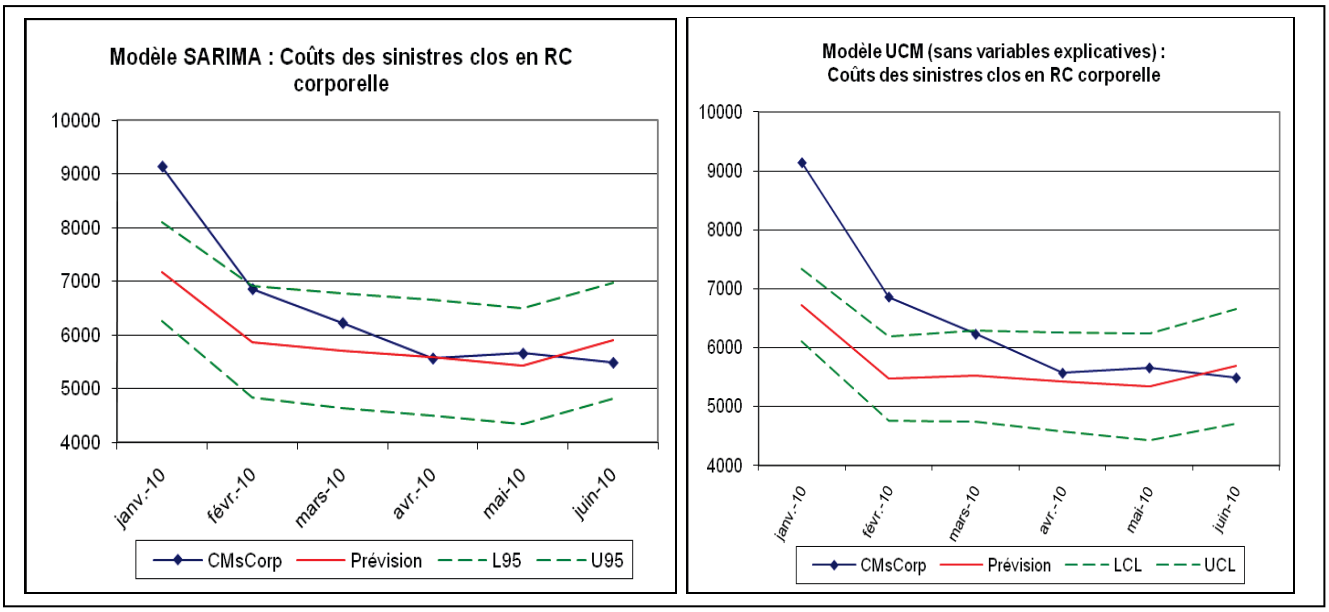
Pour la série de coûts de sinistres clos :

Les modèles que nous avons obtenus n'intègrent aucune des variables dont nous disposions pour cette étude. Il s'agit du modèle $SARIMA_{12}[(1,0,0)(0,0,0)]$ et du modèle UCM saturé. La comparaison entre les différents modèles est résumée ci-dessous :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(2) $SARIMA_{12}[(1,0,0)(0,0,0)]$	468,5	5,1%	2621,2	43,7%
(1) UCM saturé	328,6	4,3%	827,1	12,9%

On choisirait avec ce tableau le modèle UCM saturé.

Sur le premier semestre 2010, on obtient les prévisions ci-dessous :



Nous retenons le modèle UCM (il a les intervalles de confiance les moins larges sur le premier semestre). On constate cependant une forte hausse début 2010, non prévu par les deux modèles et qui est dû au nombre très faible de sinistres (– 11% par rapport au mois de janvier 2009) qui ont pourtant eu un coût élevé (23% de plus que le coût moyen du mois de janvier 2009).

De cette analyse des coûts clos en RC corporelle ressort la difficulté de modéliser ce type de coût. La RC corporelle est en effet une garantie pour laquelle les coûts varient énormément. Là où il s'agissait de réparer un bien, il va falloir indemniser une victime. Les sinistres corporels sont donc très sensibles aux évolutions de l'environnement juridique. A titre d'exemple, les assureurs soutiennent l'utilisation de la nomenclature de Dintilhac. Celle-ci permet de qualifier l'ensemble des postes de préjudices d'une victime d'un accident. Elle permet une meilleure évaluation des coûts et une plus grande transparence dans le traitement des victimes. Cependant, les assureurs constatent une évolution des tribunaux dans l'usage de cette nomenclature. Là où l'indemnisation était évaluée globalement, les postes de préjudices sont maintenant détaillés, ce qui se traduit par une inflation du coût total, chaque poste étant désormais indemnisé. Et les décisions de justice peuvent énormément varier d'une cour de justice à une autre.

Aussi, la multiplication des postes de préjudice n'est pas la seule cause de l'inflation du coût des sinistres corporels. Il faut également évoquer l'augmentation des coûts de chaque poste.

Pour citer les coûts liés à la tierce personne, en passant souvent d'un accompagnement de 8 heures par jour à du 24 heures sur 24, ces coûts deviennent extraordinairement lourds et connaissent une forte progression.

C'est pour ces diverses raisons qu'il nous est difficile de prévoir avec plus de précision l'évolution des coûts de sinistres clos en RC corporelle.

VI.2. La garantie RC matérielle

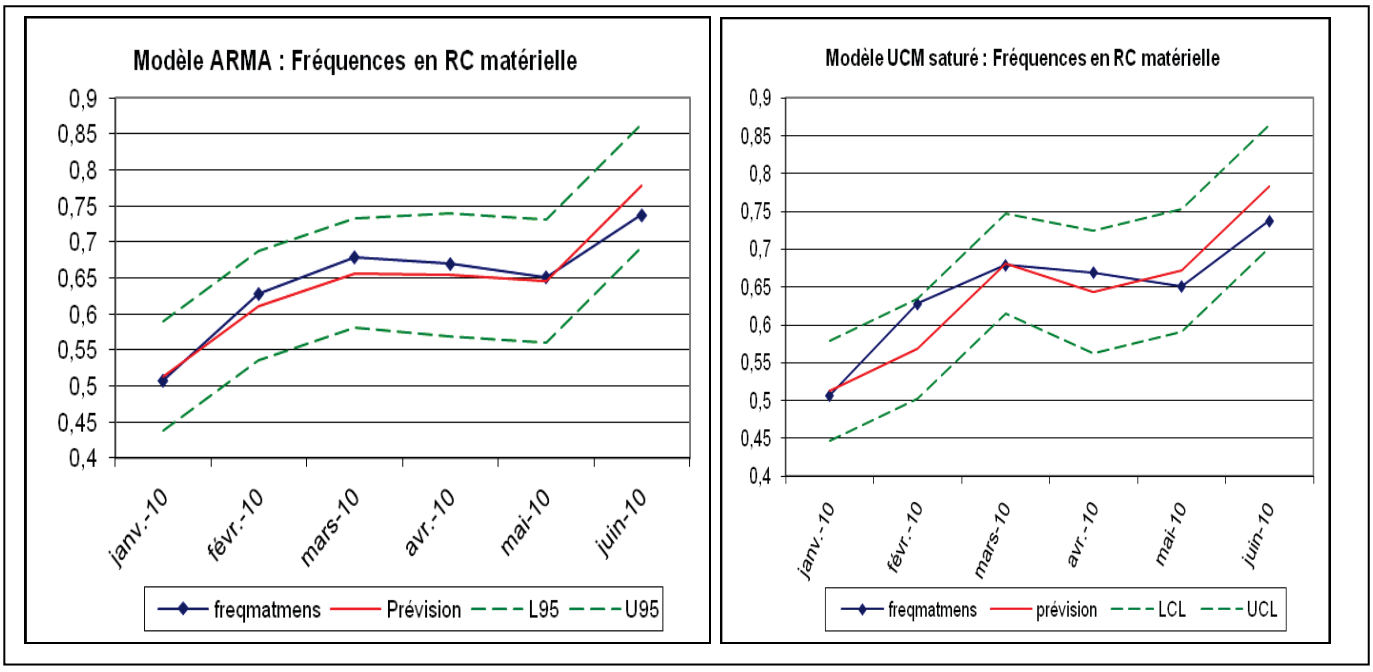
Pour la série de fréquences mensuelles :

Les modèles que nous avions retenus étaient le modèle $SARIMA_{12}[(3,0,0)(0,0,0)]$, le modèle de régression avec erreurs ARMA et le modèles UCM saturé.

Ces modèles nous donnent les critères de performance suivants :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(3) $SARIMA_{12}[(3,0,0)(0,0,0)]$	0,0387	4,8%	0,084	11,0%
(1) Régression avec erreurs ARMA	0,0194	2,4%	0,043	6,5%
(2) UCM saturé	0,0371	4,6%	0 ,088	11,6%

Au vu de ce tableau, le modèle qu'on choisirait serait le modèle de régression avec erreurs ARMA. On réalise tout de même les prévisions pour le premier semestre 2010. Les comparaisons obtenues sur le 1er semestre 2010 sont représentées ci-dessous :



Les prévisions pour le modèle de régression avec erreurs ARMA n'ont pu être réalisées sur le premier semestre 2010, car nous ne disposions pas encore des variables explicatives.

Se pose ici, de manière évidente, le problème de l'intégration de variables explicatives dans les modèles. En effet, pour faire nos prévisions, nous devons disposer de l'information sur les variables explicatives ou du moins d'une estimation. Ce qui rend encore plus complexes les prévisions, car on utilise des approximations pour élaborer nos modèles.

Le modèle idéal serait la régression avec erreurs ARMA mais à défaut de ce modèle, on peut utiliser, en deuxième option, un modèle UCM sans variables explicatives.

Pour la série de coûts de sinistres clos :

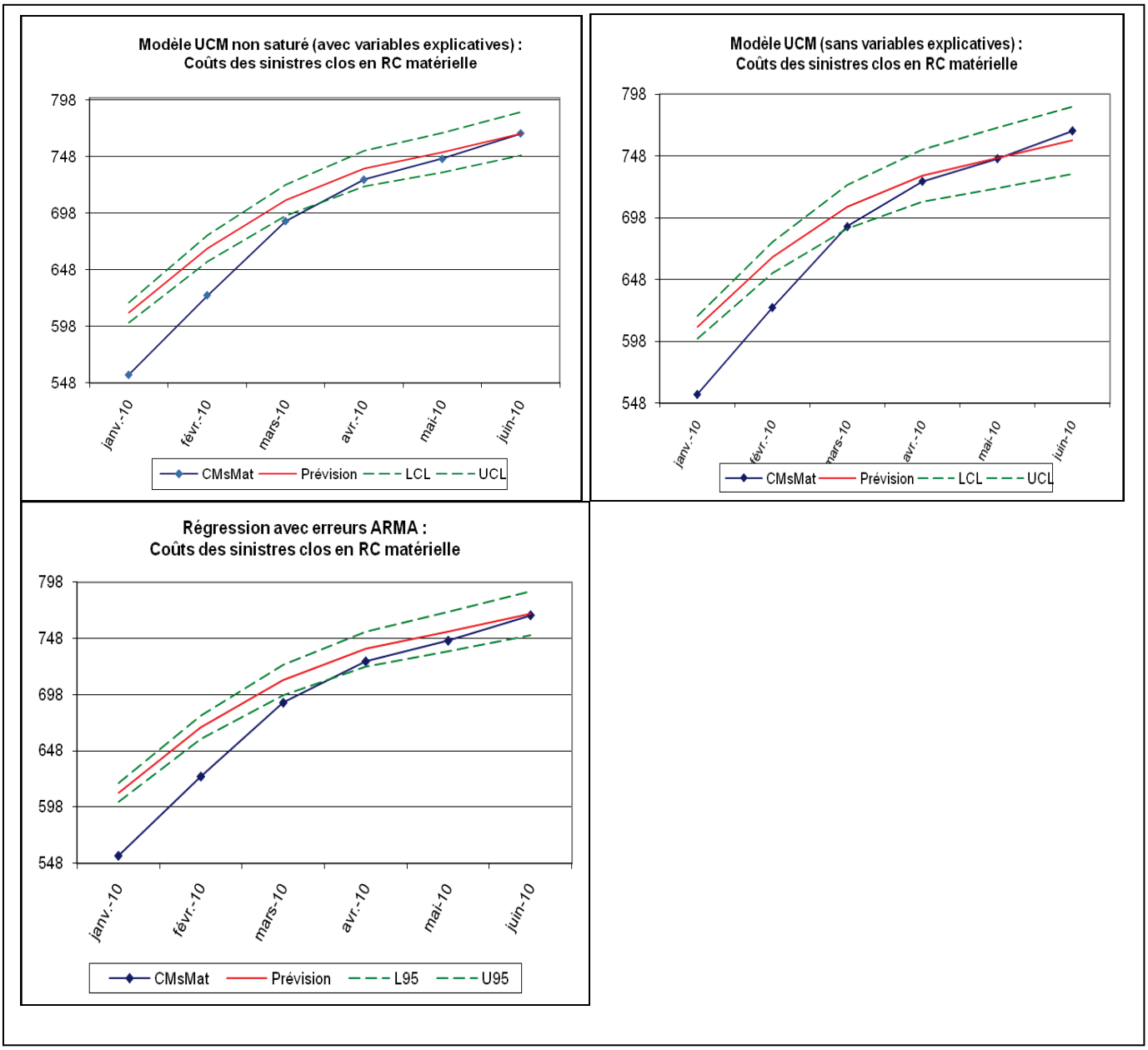
Les modèles que nous avons obtenus sont le modèle $SARIMA_{12}[(1,1,0)(0,0,0)]$, le modèle de régression avec erreurs ARMA et les modèle UCM (avec et sans variables explicatives). Mais le modèle SARIMA n'a pas franchi l'étape de la validation. Donc nous n'avons gardé que le modèle de régression avec erreurs ARMA et les modèles UCM (non saturé et saturé).

D'un point de vue « critères », nous avons obtenu les résultats suivants :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(1) Régression avec erreurs ARMA	3,9	0,4%	11,2	1,8%
(3) Modèle UCM saturé	5,5	0,6%	18,8	2,9%
(2) Modèle UCM non saturé	4,9	0,5%	13,1	2,1%

Le meilleur modèle, avec les erreurs les plus faibles, est le modèle de régression avec erreurs ARMA.

D'un point de vue « prévision », la comparaison sur le premier semestre 2010 est illustrée ci-après :



Le modèle que nous retenons est le modèle de régression avec erreurs ARMA.

On constate cependant que tous nos modèles ont du mal à prévoir les trois premiers mois de l'année : ils avaient prévu des coûts moyens plus élevés que ceux réellement observés sur ces mois. Les coûts de sinistres clos en RC matérielle ne dépendant pas uniquement des prix des pièces détachées et accessoires. Les dédommagements de sinistres matériels vont permettre de régler des travaux de réparations pour reconstituer le véhicule. Les coûts de ces sinistres sont donc liés au prix de la main-d'œuvre, des véhicules ou de leurs pièces détachées, etc. Ce qui explique pourquoi notre modèle, même en ayant intégré l'indice des pièces détachées et accessoires n'arrive pas à être « bon » sur tous les mois : il nous manque des variables explicatives complémentaires.

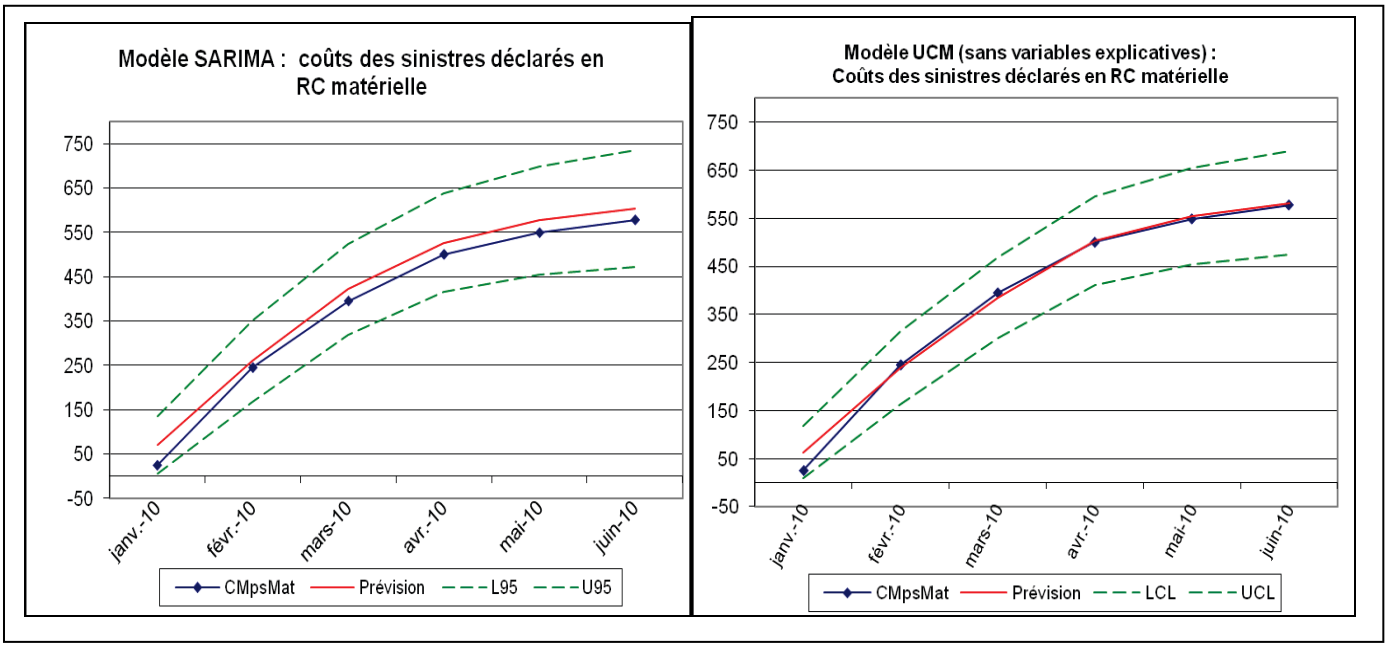
Pour la série de coûts de sinistres déclarés :

La garantie RC matérielle étant à déroulement court, et au vu des résultats obtenus avec les coûts de sinistres clos, nous mentionnons ici les résultats obtenus sur les sinistres déclarés. En effet, nous sommes arrivés à des modèles très intéressants, sans variables explicatives comparés ci-dessous :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(2) $SARIMA_{12}[(1,1,0)(0,0,0)]$	32,8	19,6%	177,7	381,1%
(1) Modèle UCM saturé	32,2	22,1%	161	243,6%

On observe des erreurs très importantes à cause du mois de février 2005, qui a eu un coût moyen relativement faible, par rapport aux mois de février des années précédentes.

Les prévisions sur le premier semestre 2010 sont représentées ci-dessous :



On constate que les prévisions de ces modèles sont beaucoup plus près du réel que les prévisions sur les coûts de sinistres clos. Ils intègrent le retard d'ordre 2 de la variable coût.

Au vu des résultats obtenus sur le premier semestre 2010, nous avons opté pour la modélisation des coûts de sinistres déclarés en RC matérielle, avec le modèle UCM sans variables explicatives.

VI.3. La garantie Dommages tous accidents

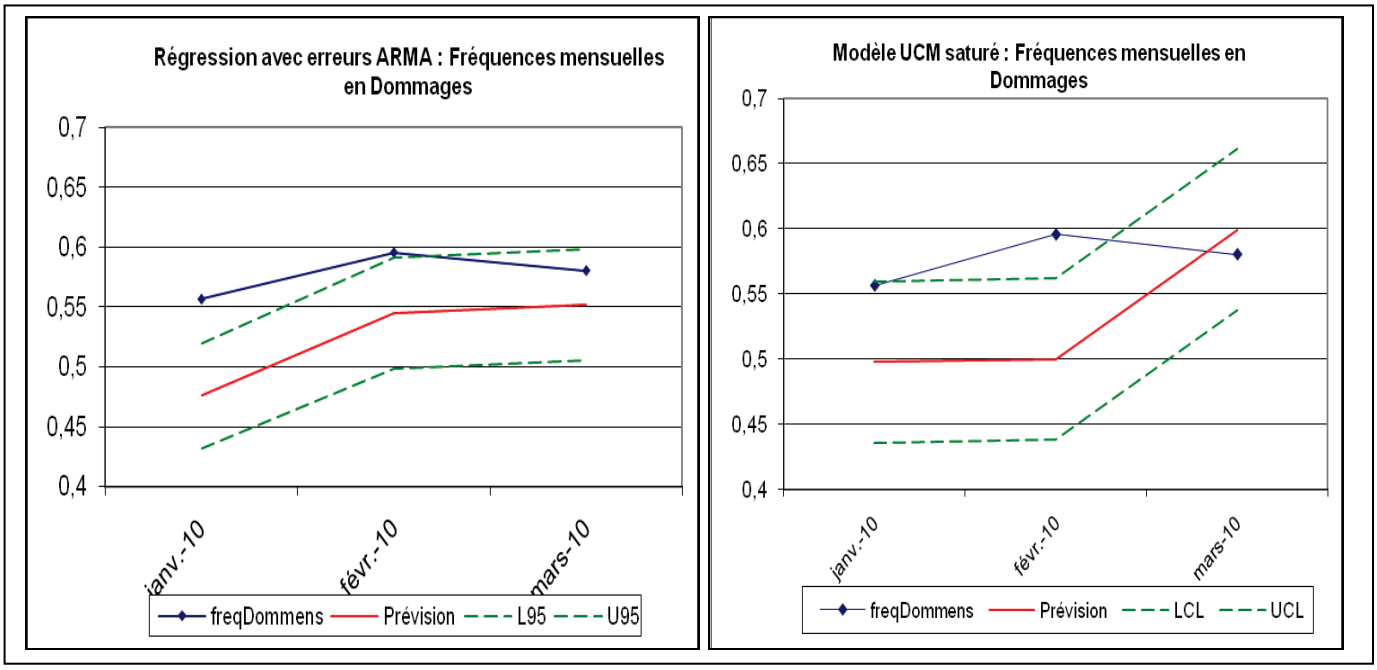
Pour la série de fréquences mensuelles :

Les modèles présélectionnés pour cette série étaient le modèle de régression avec erreurs ARMA, le modèle UCM saturé et le modèle UCM non saturé. En comparant les différents critères du tableau, le modèle qu'on serait tenté de retenir est le modèle de régression avec erreurs ARMA.

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(1) Régression avec erreurs ARMA	0,0206	2,8%	0,05	11,7%
(2) UCM saturé	0,034	5%	0,09	12,7%

Nous n'avons pu obtenir les observations des variables explicatives du modèle de régression avec erreurs ARMA et du modèle UCM non saturé que sur les 3 premiers mois de l'année ; c'est pour cette raison que nos prévisions s'arrêtent au mois de mars 2010 avec ces modèles. En comparant les prévisions sur l'année 2010, le modèle qu'on conserve est le modèle de régression avec erreurs ARMA. En deuxième option, on pourra opter pour un modèle UCM saturé.

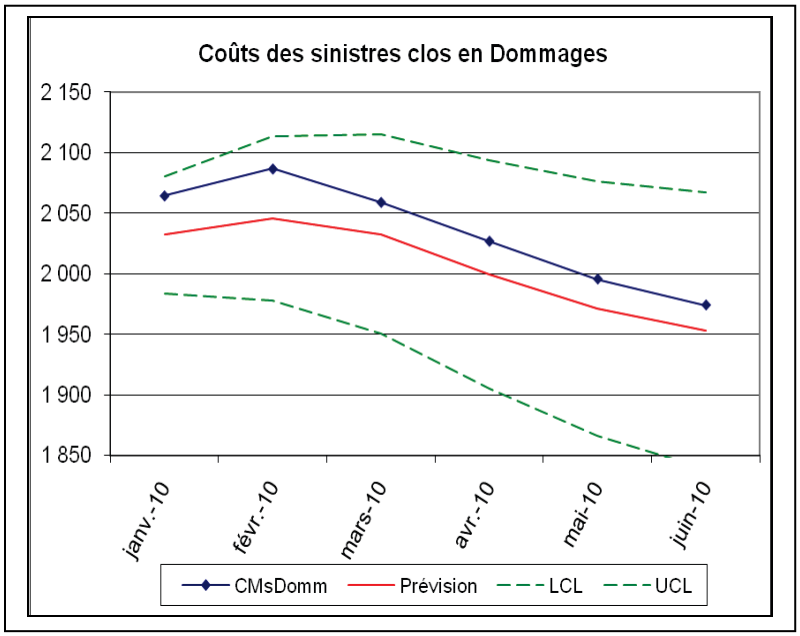
Les prévisions obtenues pour le début 2010 sont représentées ci-dessous :



Pour la série de coûts de sinistres clos :

Le modèle que nous avons retenu est le modèle UCM saturé (sans variables explicatives). En effet, la modélisation ARMA ne nous permettait pas d'obtenir de modèles : une fois retirées la tendance et la saisonnalité, la série devenait un bruit blanc.

Les prévisions pour le premier semestre 2010 sont représentées ci-dessous :



On constate que nos prévisions restent dans l'intervalle de confiance, même si les intervalles de confiance s'élargissent et qu'on reste en-dessous du réel observé.

VI.4. La garantie Bris de glace

Pour la série de fréquences mensuelles :

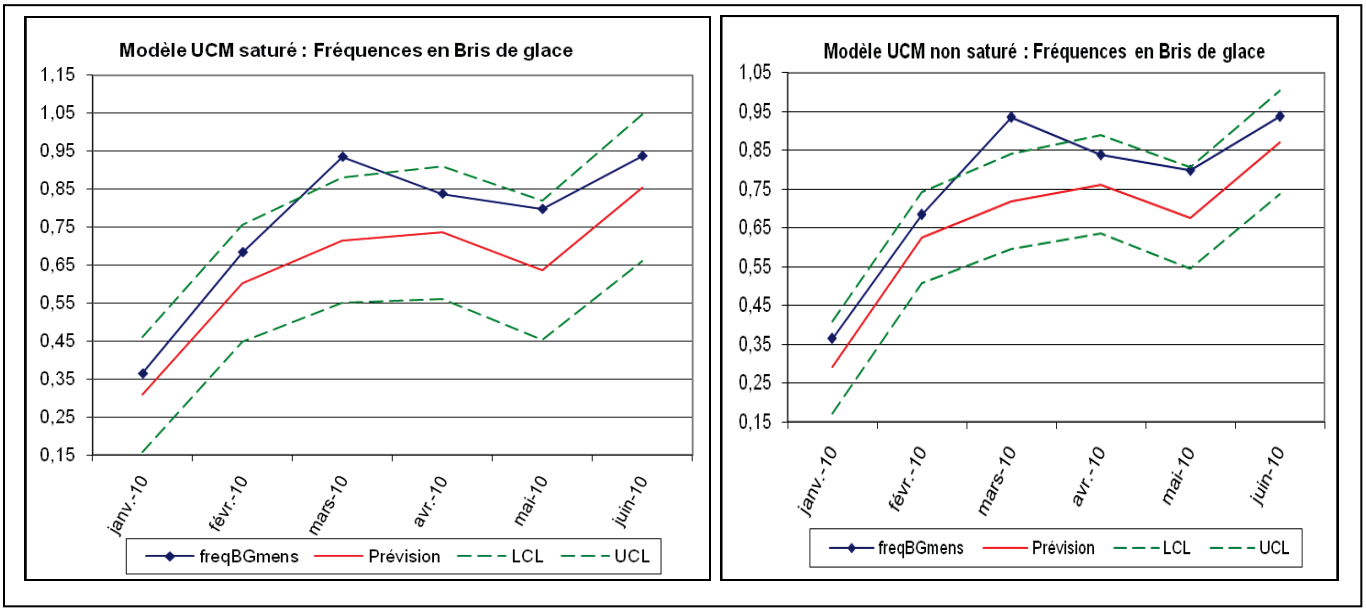
Nous avions réalisé quatre modèles sur cette série : le modèle $SARIMA_{12}[(1,0,1)(0,0,0)]$, le modèle de régression avec erreurs ARMA, le modèle UCM saturé et le modèle UCM non saturé. Le modèle SARIMA n'a pas réussi l'épreuve de la validation donc nous l'excluons de cette partie.

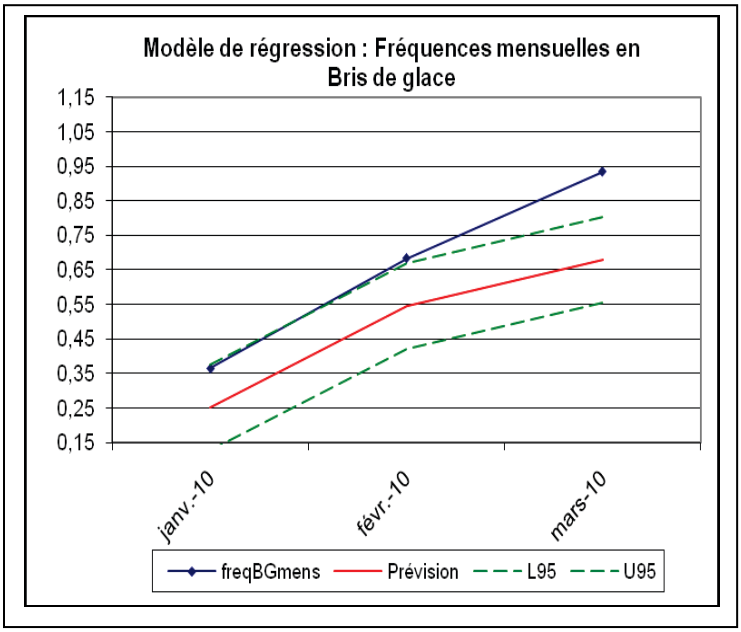
Les critères de performance de ces différents modèles sont résumés dans le tableau ci-dessous :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(1) Régression avec erreurs ARMA	0,060	6,89%	0,24	49,24%
(3) Modèle UCM saturé	0,078	9,78%	0,20	37,3%
(2) Modèle UCM non saturé	0,063	7,37%	0,17	37,0%

Avec ce tableau, on serait tenté de choisir le modèle de régression avec erreurs ARMA. On a une erreur moyenne de plus de 5%, quelque soit le modèle choisi.

Les prévisions pour l'année 2010 pour chacun des modèles sont les suivantes :

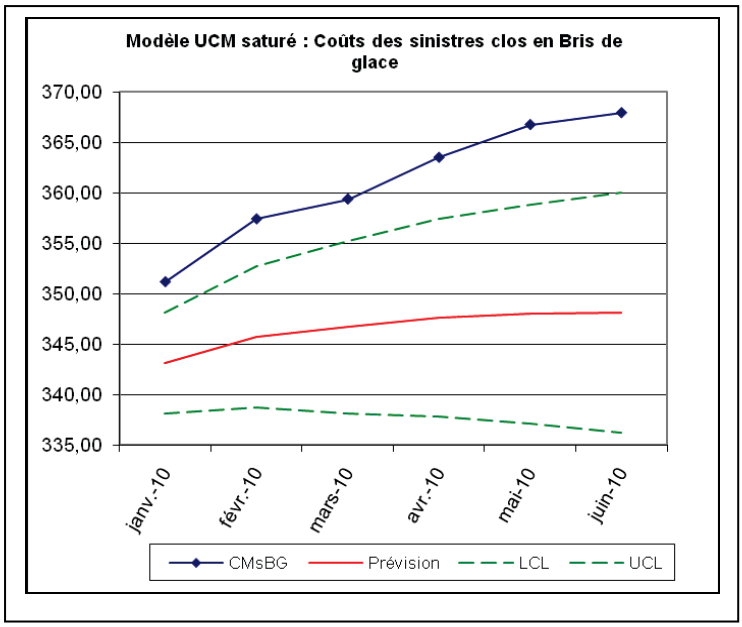




Les prévisions pour le modèle de régression n'ont pu être réalisées que sur les 3 premiers mois : nous ne disposions pas des valeurs de toutes les variables explicatives au-delà de ces 3 mois. Au vu de ces graphiques, nous avons opté pour le modèle UCM non saturé, avec comme variable explicative la consommation de carburants.

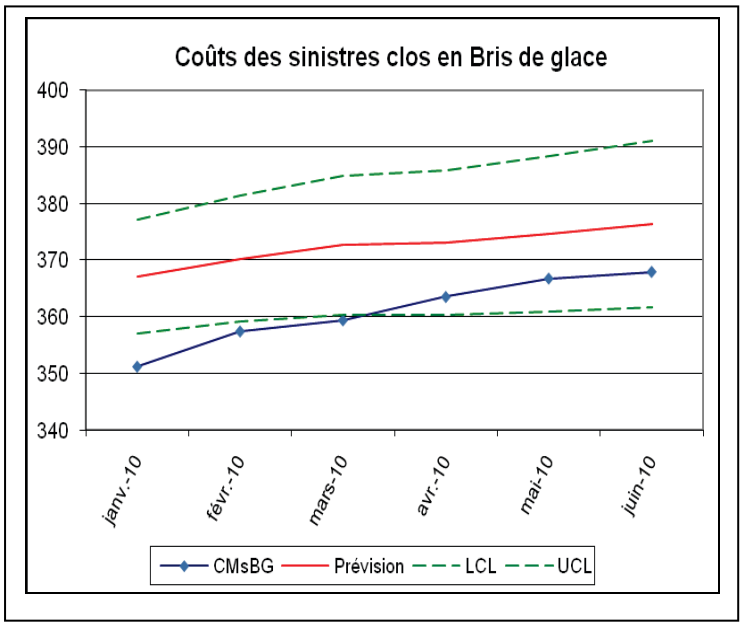
Pour la série de coûts de sinistres déclarés :

Nous n'avions que le modèle UCM saturé sur cette série. Malgré que ce modèle ait passé l'étape de la validation, les prévisions obtenues pour le premier semestre 2010 ne sont pas satisfaisantes :



Pour la série de coûts de sinistres clos :

Le modèle que nous avons obtenu intègre l'indice des prix à la consommation des véhicules personnels.



On retiendra donc uniquement le modèle UCM non saturé sur la série des coûts de sinistres clos en Bris de glace.

VI.5. La garantie Vol partiel

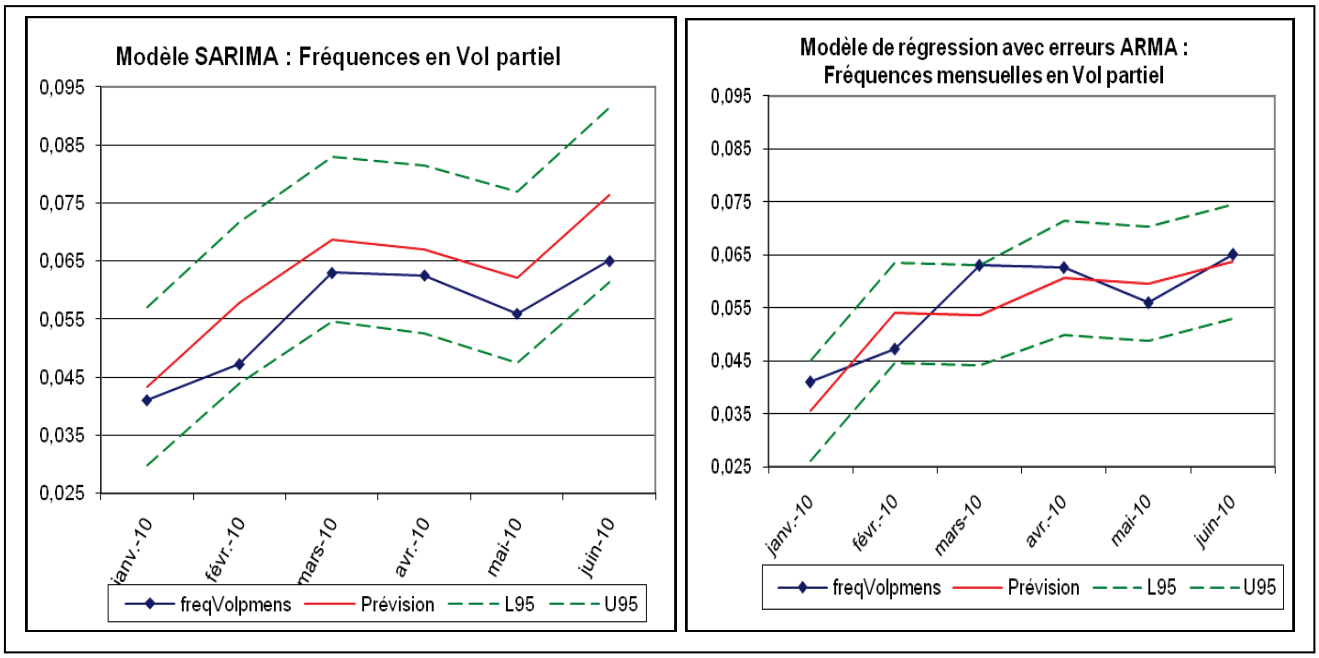
Pour la série de fréquences mensuelles :

Les modèles que nous avions retenus après les phases de réalisation et de validation était le modèle $SARIMA_{12}[(1,1,1)(0,0,0)]$, le modèle de régression avec erreurs ARMA, le modèle UCM saturé et le modèle UCM non saturé. Le modèle UCM non saturé n'est pas stable dans le temps, donc nous le retirons de nos modèles potentiels.

La comparaison des modèles selon les différents critères est la suivante :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(2) $SARIMA_{12}[(1,1,1)(0,0,0)]$	0,007	7,7%	0,017	21,9%
(1) Régression avec erreurs ARMA	0,004	5,0%	0,013	18,9%

Les prévisions sur le premier semestre 2010 sont représentées ci-dessous :

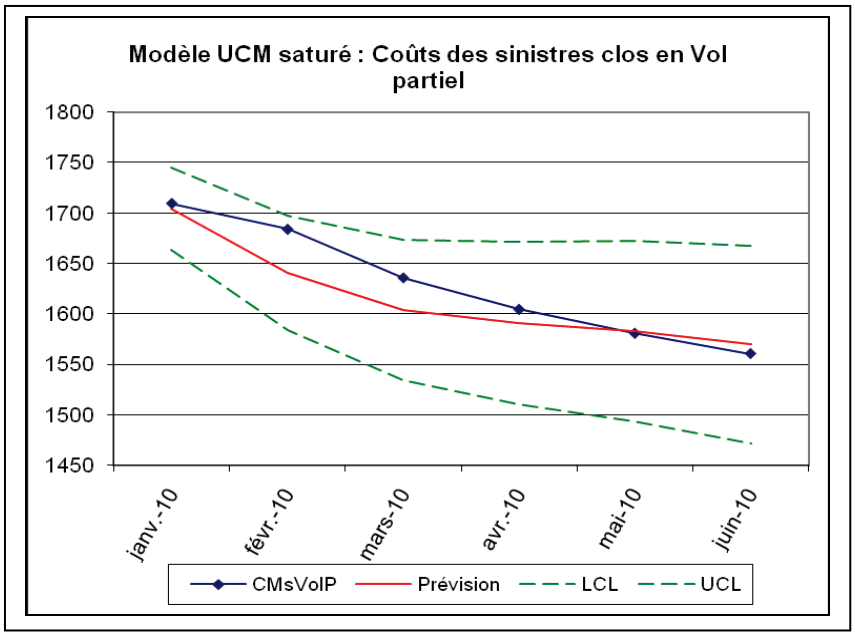


On constate que le modèle SARIMA nous donne de meilleures prévisions sur le premier semestre 2010, malgré l'erreur moyenne plus importante, c'est donc ce modèle que nous garderons.

Pour la série de coûts de sinistres clos :

Cette série n'avait pas pu être étudiée dans la partie « modélisation ARMA », car elle devenait un bruit blanc après le retrait de la tendance et de la saisonnalité. Avec la modélisation UCM, nous avons obtenu le modèle suivant :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
Modèle UCM saturé	22,04	1,0%	113,36	6,8%



Nous constatons que les données observées sur le premier semestre 2010 restent dans la bande de confiance de notre modèle.

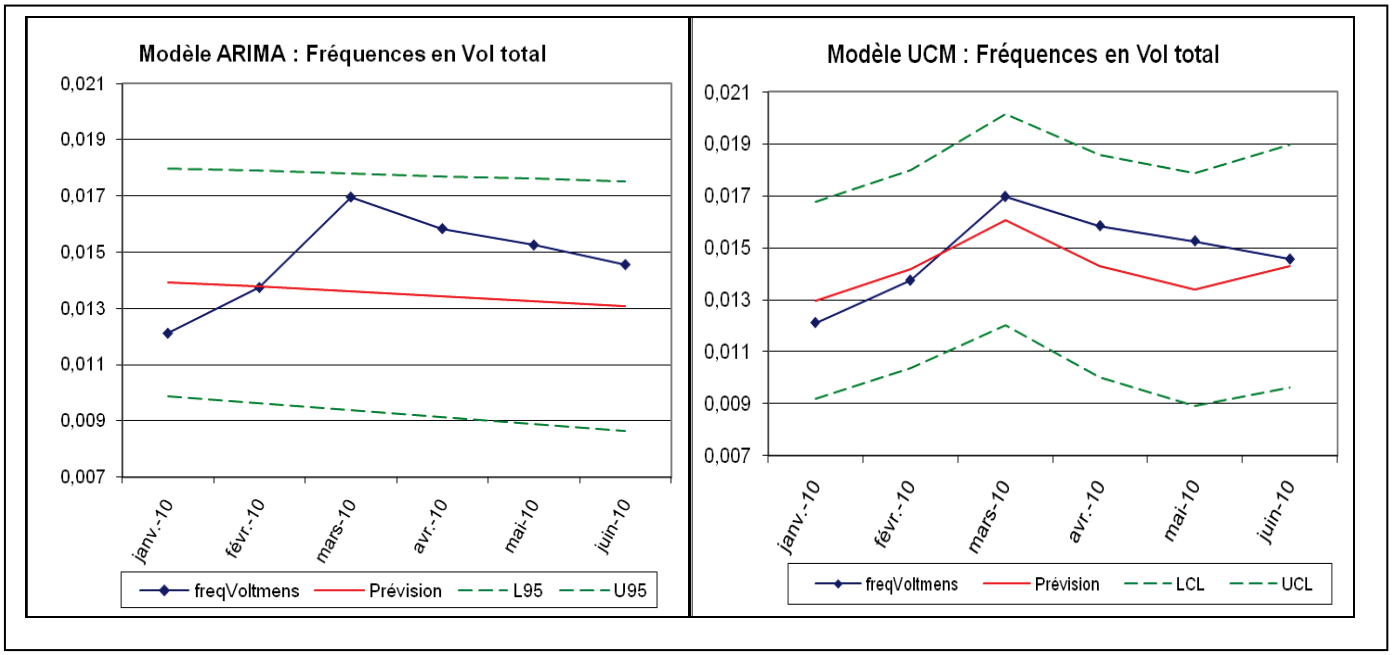
VI.6. La garantie Vol total

Pour la série de fréquences mensuelles :

Les modèles que nous avions retenus après les phases de réalisation et de validation était le modèle ARIMA(0,1,1), le modèle de régression avec erreurs ARMA, et le modèle UCM saturé. La comparaison des modèles selon les différents critères est la suivante :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(2) ARIMA(0,1,1)	0,0020	8,3%	0,007	30,9%
(1) Régression avec erreurs ARMA	0,0014	6,3%	0,004	19,1%
(3) Modèle UCM saturé	0,0020	8,5%	0,005	24,0%

Les prévisions pour le premier semestre 2010 sont représentées ci-dessous :



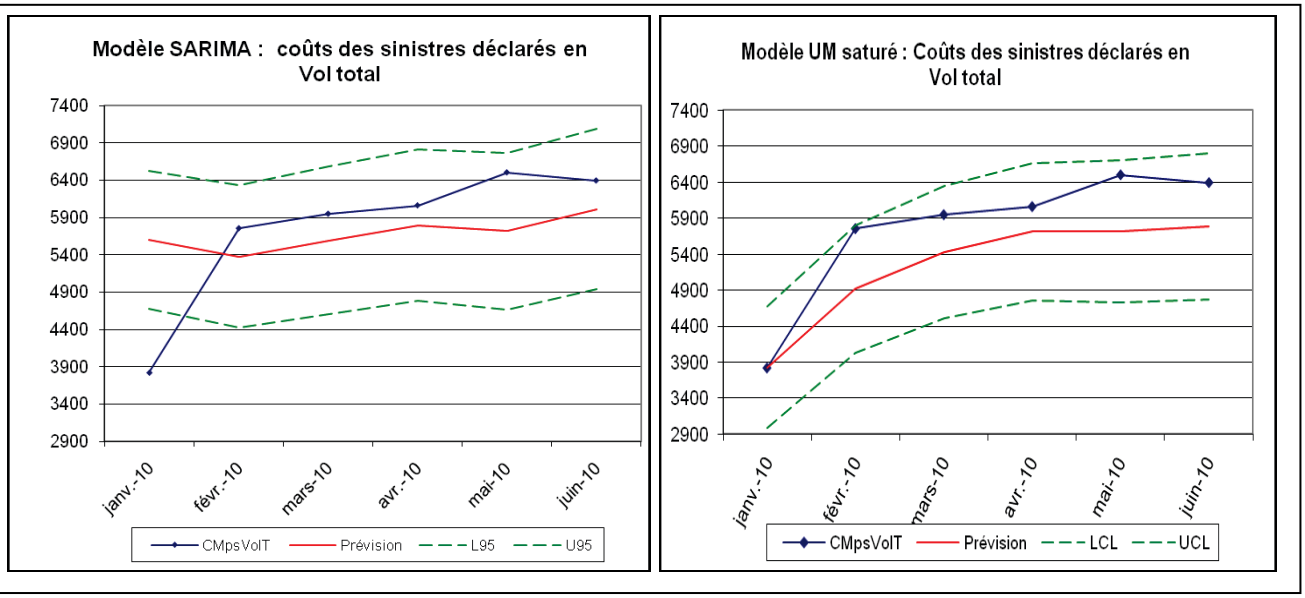
Pour le modèle de régression avec erreurs ARMA, nous ne disposions pas de toutes les variables explicatives pour faire les prévisions sur le premier semestre de 2010.
Le modèle UCM saturé est celui que nous retenons.

Pour la série de coûts de sinistres déclarés :

Les modèles que nous avions retenus après les phases de réalisation et de validation était le modèle $SARIMA_{12}[(0,1,1)(0,0,0)]$ et le modèle UCM saturé. La comparaison des différents modèles est résumée ci-dessous :

Critère / Modèle	RMSE	MAPE	MPE1	MPE2
(1)$SARIMA_{12}[(0,1,1)(0,0,0)]$	467,40	6,7%	2360,38	63,4%
(2) Modèle UCM saturé	462,74	7,7%	2074,03	85,2%

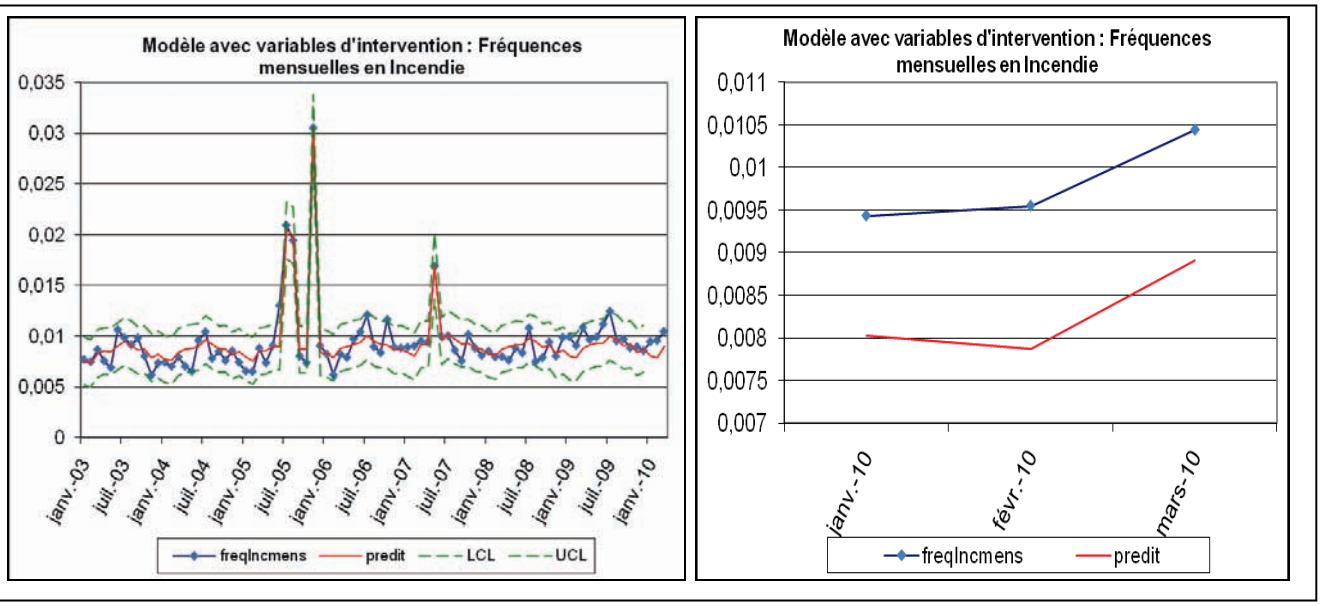
En considérant l'erreur moyenne MAPE, on choisirait le modèle SARIMA. Mais en observant les prévisions sur le premier semestre 2010, nous avons opté pour le modèle UCM saturé.



VI.7. La garantie Incendie

Pour la série de fréquences mensuelles :

La modélisation de la série Incendie est assez particulière. Nous n'avions retenu qu'une approche de type régression linéaire, avec deux méthodes : la première qui « excluait » les valeurs atypiques de la modélisation (elle en tenait compte mais les modélisaient en les remplaçant par des estimations robustes) et la deuxième qui les considérait comme des évènements rares (introduction de variables d'intervention). C'est ce deuxième modèle que nous avons retenu car il nous donnait une erreur moyenne plus faible que le premier et un plus grand R^2. Les prévisions obtenues pour l'année 2010 comparées au réel 2010 sont illustrées ci-dessous :



Les fréquences en Incendie fluctuent autour d'une moyenne de 0,0094.

CONCLUSION

Au terme de cette analyse, en considérant nos séries de fréquences et de coûts mensuels comme des séries temporelles, les résultats que nous avons obtenus se résument ainsi :

Les séries de fréquences nous donnent des modèles où la saisonnalité intervient, quelque soit la modélisation utilisée. De plus, les garanties RC corporelle et Bris de glace intègrent la consommation de carburants dans leur modélisation : plus les conducteurs roulent et plus on aura de sinistralité sur ces deux garanties. La série de fréquences mensuelles en incendie, du fait de la présence de valeurs extrêmes, ne donnait pas de résultat satisfaisant avec les modèles que nous utilisions. Les méthodes de régression robuste nous ont pourtant permis d'obtenir un modèle insensible à ces observations extrêmes. Mais pour réaliser ce modèle, nous avons pris comme hypothèse que les évènements qui avaient générés ces observations extrêmes (émeutes dans les banlieues en 2005, élections présidentielles agitées en 2007) étaient des faits ponctuels et ne se reproduiraient pas. Ce qui est une hypothèse très forte et dont va dépendre la stabilité de notre modèle au cours du temps. Il est ressorti de cette étude que les éventuelles dérives dans les fréquences de sinistres en Incendie étaient liées à des évènements sociaux ; or ces évènements ne sont pas forcément quantifiables et prévisibles à l'avance.

Les prévisions sur les séries de coûts sont celles qui ont nécessité le plus d'attention :

- Sur la RC corporelle, branche à déroulement long, les variables explicatives dont nous disposions ne nous permettaient pas d'expliquer les coûts de sinistres clos. En effet, pour évaluer les coûts de cette branche, il faudrait prendre en compte l'évolution du système juridique en termes d'indemnisation ; or cette évolution n'est pas uniforme à l'échelle nationale : l'indemnisation des différents postes de préjudice reste assez subjective d'une cour de justice à une autre et ces différents postes connaissent une inflation qui n'est, à notre connaissance, pas encore mesurée et encore moins prévisible.

- Sur la RC matérielle, nous avons pu obtenir un modèle qui prenait en compte l'inflation avec l'indice des prix à la consommation sur la série des coûts de sinistres clos. Toutefois, la modélisation des coûts de sinistres déclarés, sans variables explicatives, nous donne d'excellentes prévisions.

- Sur la garantie Dommages, c'est la série des coûts de sinistres clos que nous avons modélisé. Le modèle que nous avons obtenu ne prend en compte aucune des variables explicatives de notre sélection. Les prévisions sont néanmoins très bonnes.

- Sur la garantie Bris de glace, nous retrouvons l'influence de l'indice des prix à la consommation sur les véhicules personnels

- Sur les garanties Vol partiel et Vol total, nous avons obtenu des modèles sans variables explicatives, avec toutefois des prévisions qui cadraient avec le réel observé. Nous n'avons pas rajouté de variables explicatives liées aux comportements humains (délinquance, fraude etc.) car l'aspect que nous voulions faire ressortir était surtout l'influence des caractéristiques des conducteurs sur la sinistralité.

Aussi, il s'est avéré que les données portefeuille (exprimées en années police ou durée d'exposition) n'avaient pas d'impact sur l'évolution des fréquences et des coûts mensuels. En effet, ces données étant annuelles, nous avons dû les mensualiser ; mais pour rester cohérents avec l'information qu'elles apportaient, nous les avions démultipliées. Aucune variation mensuelle n'apparaissant, les données ne fluctuaient pas suffisamment pour expliquer nos séries mensuelles. Une autre approche qu'on pourrait tester serait de se restreindre uniquement aux données portefeuille des assurés ayant été sinistrés et qui possédaient la garantie cible (par exemple la garantie Dommages ou Incendie-Vol). Par manque de temps, nous n'avons pas pu poursuivre sur cette piste.

Les données météorologiques dont nous disposons ne mesurent pas exactement l'intensité des phénomènes climatiques ponctuels qui ont touché la France ces dernières années. Il n'existe actuellement pas de systèmes de recensement de l'intensité de toutes les catastrophes naturelles à l'échelle nationale. En utilisant les moyennes mensuelles, on ne retrouve pas l'information sur la sévérité d'une variation ponctuelle de la donnée météorologique.

Enfin, le fait d'intégrer des variables explicatives dans les modèles leur rajoute de la complexité. Avant de faire des prévisions de notre série de fréquence ou de coût, il nous faudra faire une estimation des variables explicatives qui entrent dans le modèle. On introduit un aléa supplémentaire, ce qui diminue la précision de nos prévisions. Mais on peut également réaliser un certain nombre de scénarios à partir de différentes valeurs probables des variables explicatives. L'approche globale du risque qu'on a est alors plus précise.

BIBLIOGRAPHIE

Mémoires & Thèses:

Bergel-Hayat R., *La prise en compte de variables explicatives dans les modèles de séries temporelles – Applications à la demande de transport et au risque routier*. Thèse Université Paris-Est, Ecole Doctorale ICMS 2008.

Mbengue Y., *Modélisation de la sinistralité par garantie en Auto-particuliers 4 roues*. Stage AXA DT IARD service Auto particuliers 2009.

Articles :

Chen C., *Robust Regression and Outlier Detection with the ROBUSTREG Procedure.* SAS Institute Inc., Cary, NC. SUGI27 Statistics and Data Analysis. http://www2.sas.com/proceedings/sugi27/p265-27.pdf

Cornu J., Gatterer B., *Evolution récente de l'assurance de biens des particuliers*. Risques, Les cahiers de l'assurance N° 80, Décembre 2009. http://www.ffsa.fr/webffsa/risques.nsf/html/Risques_80_0019.htm

De Peretti J., *Les mouvements de prix et l'assurance IARD*. Risques, Les cahiers de l'assurance N° 80, Décembre 2009.

Gharibvand L., *Using Unobserved Components Model (UCM) for a Stock Price Fluctuation*. University of California, Riverside. http://www.wuss.org/proceedings08/08WUSS%20Proceedings/papers/anl/anl12.pdf

Kohn R., Shively S.T., Ansley F.C., *Algorithm AS 279: Computing p-Values for the Generalized Durbin-Watson Statistic and Residual Autocorrelations in Regression.* Journal of the Royal Statistical Society. Series C (Applied Statistics) Vol. 42, No. 1. http://www.jstor.org/stable/2347430

Ladiray D., *Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles. Décembre 2002.* http ://www.unige.ch/ses/sococ/mirage

Lavery R., *An Animated Guide©: Proc UCM (Unobserved Components Model).* NESUG17 Analysis http://www.nesug.org/proceedings/nesug04/an/an03.pdf

Lemoine M., Pelgrin F., *Introduction aux modèles espace-état et au filtre de Kalman*. Revue de l'OFCE n° 86 Juillet 2003 http://www.ofce.sciences-po.fr/pdf/revue/8-86.pdf

Ragavan A.J., Fernandez G.C., Modeling Water Quality Trend in Long Term Time Series. University of Nevada, Reno, NV 89557. SUGI31 Statistics and Data Analysis. http://www2.sas.com/proceedings/sugi31/205-31.pdf

Selukar R., *Structural Analysis of Time Series Using the SAS/ETS® UCM Procedure*. SAS Institute Inc., Cary, NC. http://www.sascommunity.org/mwiki/images/5/5a/Ucm.pdf

Tervola J., *Robust regression analysis of registered-based sickness insurance data.* University of Helsinki, August 2010. http://vilniusworkshop2010.stat.gov.lt/Straipsniai/Tervola_J.pdf

Yaffee A.R., *Robust regression analysis: Some popular statistical package options* http://pdfcast.org/pdf/robust-regression-analysis-some-popular-statistical-package-options

Cours :

Ivers H., *Analyse d'intervention en séries chronologiques*. Université Laval, Ecole d'été 2004.

Karamé F., *Notes de cours - Atelier macroéconomique*. Université d'Evry, IUP Ingénierie Economique Statistique 2009.

Kratz M., *Cours de séries temporelles*. Institut de Statistiques de l'Université de Paris 2010.

Livres :

Brockwell P.J., Davis R.A., *Introduction to Times Series and Forecasting*. Springer Texts in Statistics, Springer. New-York, 1996.

Commandeur J.J.F., Koopman S.J., *An introduction to State Space Time Series analysis*. Oxford University Press 2007.

Droesbeke J.-J., Fichet B., Tassi P., eds (1989) *Séries chronologiques : théorie et pratique des modèles ARIMA*. Economica, Paris.

Gouriéroux C., Montfort A., *Séries temporelles et modèles dynamiques 2ème édition*. Collection ÉCONOMIE ET STATISTIQUES AVANCEES, 1995.

Harvey A.C., *Forecasting, Structural time series models and the Kalman filter*. Cambridge University Press 1989.

Harvey A.C., Koopman S.J., Sheiphard N., *State Space and unobserved component models: Theory and Applications*. Cambridge University Press 2004.

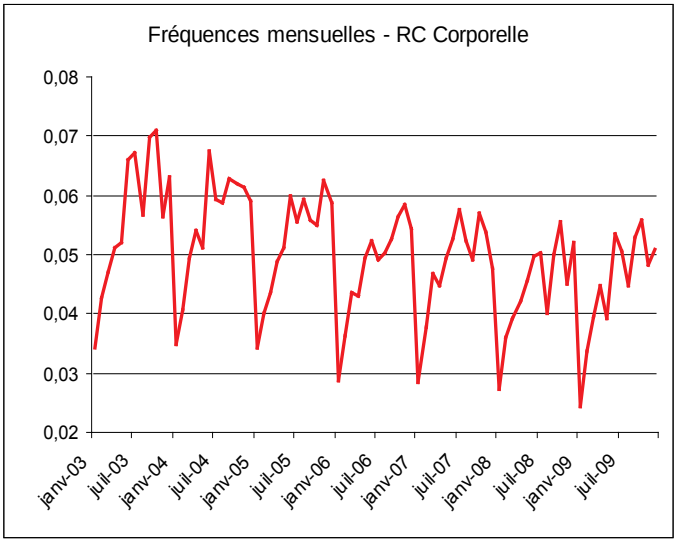
Saporta G., *Probabilités, analyse des données et statistique 2ème édition*. Editions TECHNIP 2006.

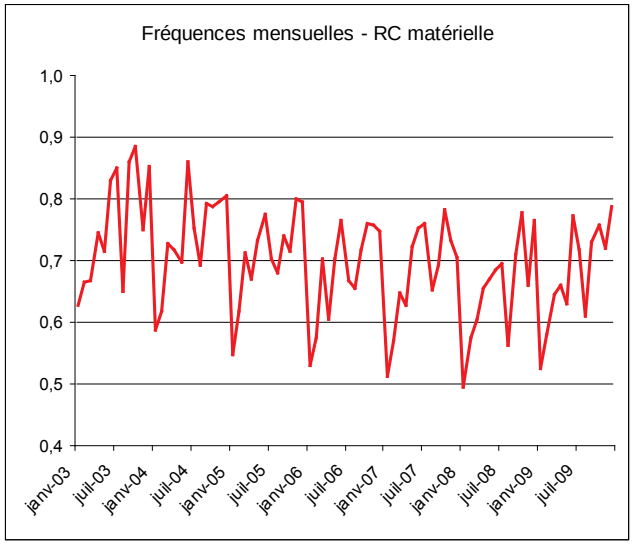
Support SAS 9.1. http://support.sas.com/documentation/onlinedoc/91pdf/index.html

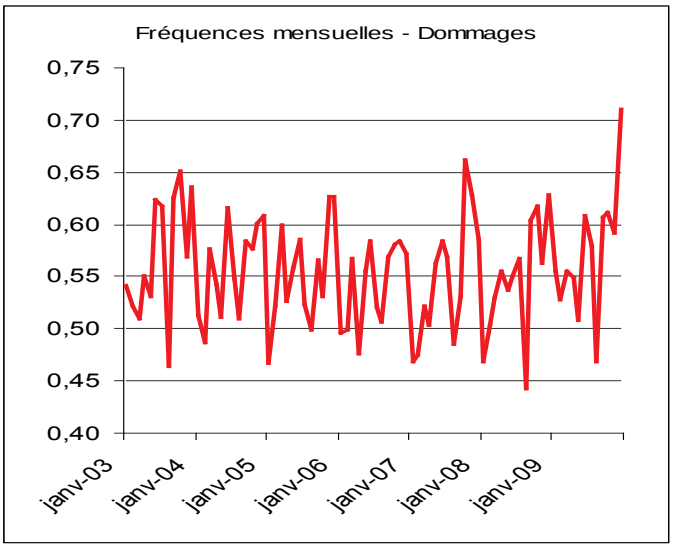
TABLE DES ILLUSTRATIONS

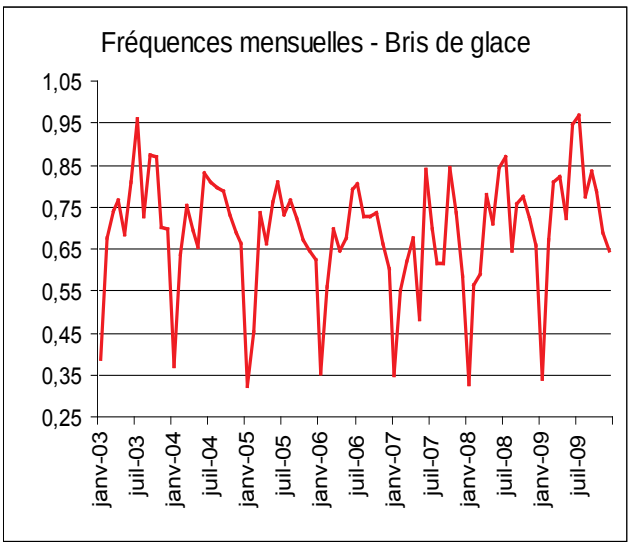
ANNEXES

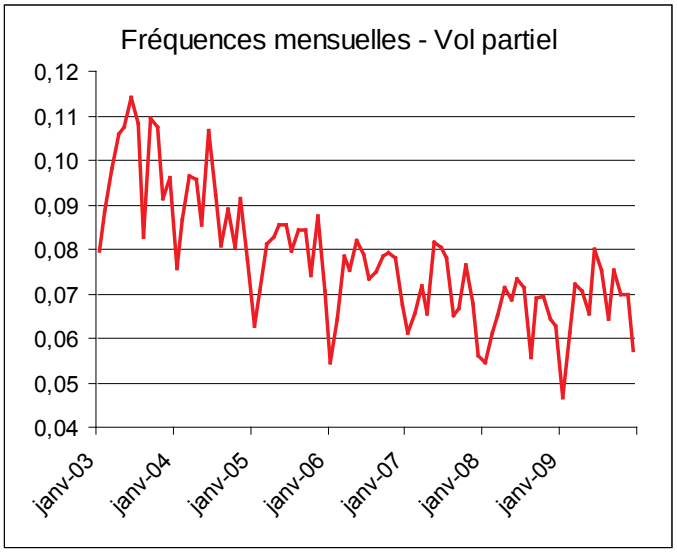
ANNEXE 1 : Graphiques des séries de fréquences et de coûts à modéliser

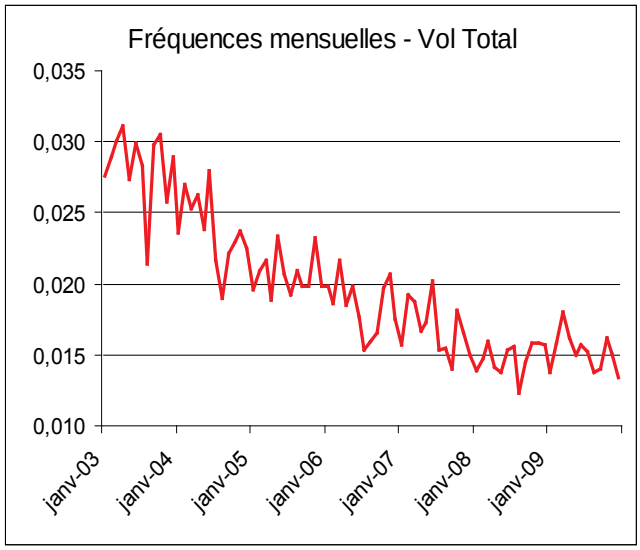


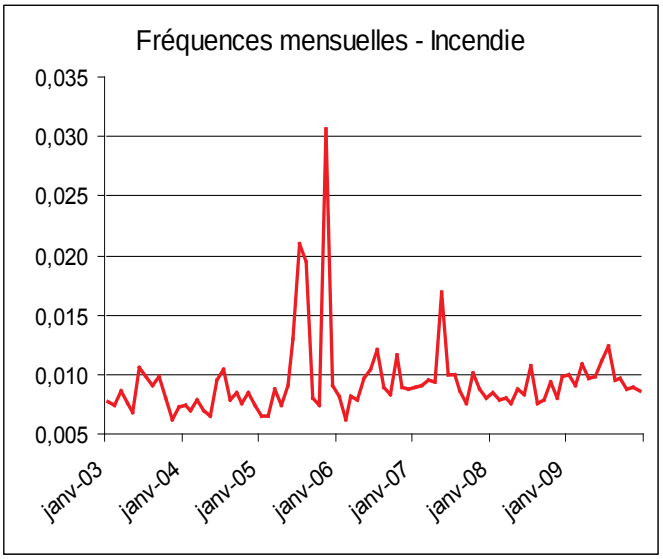


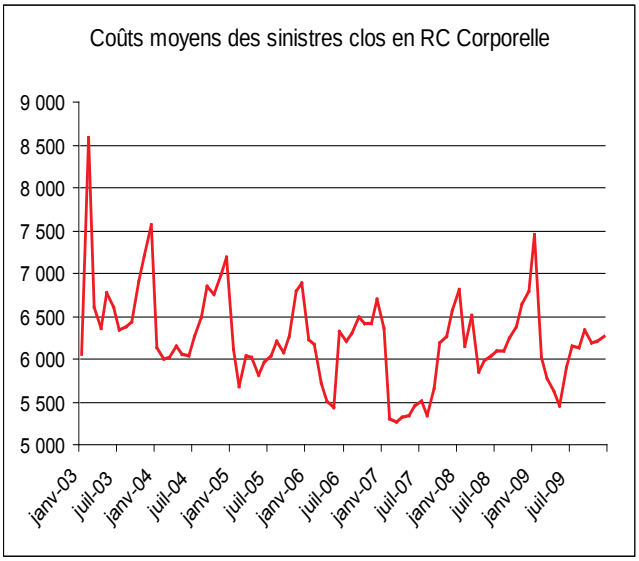


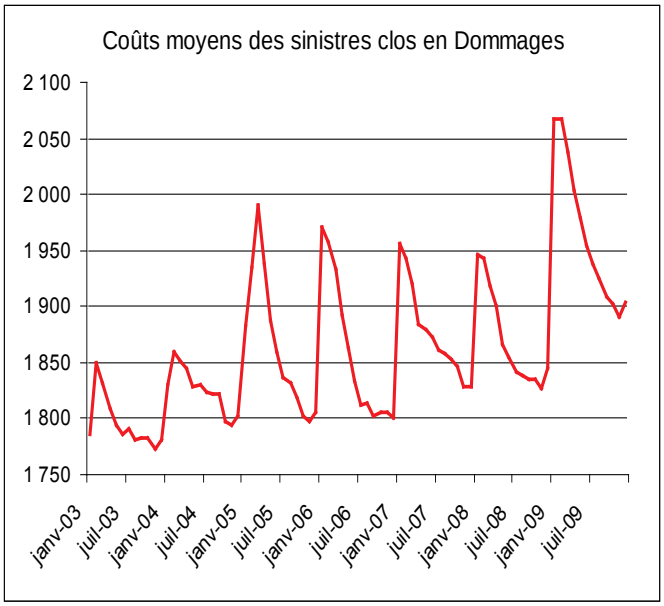


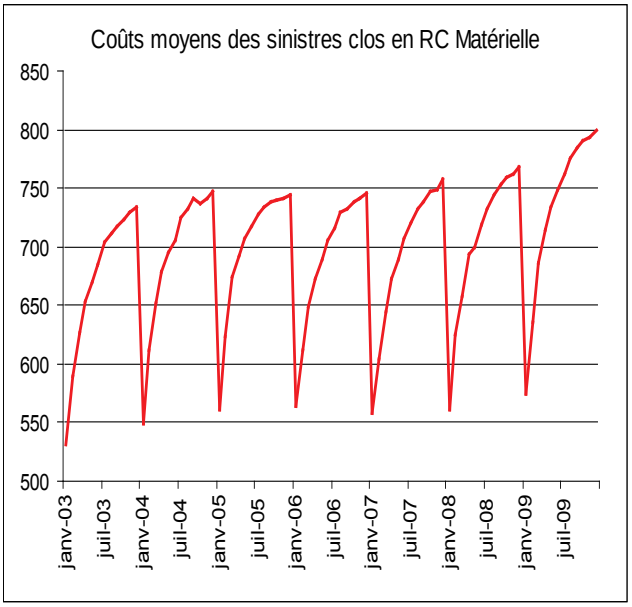


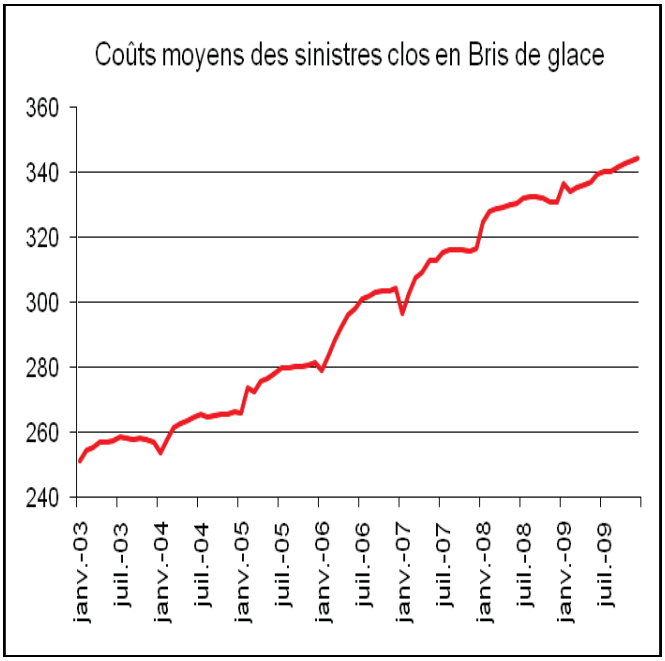


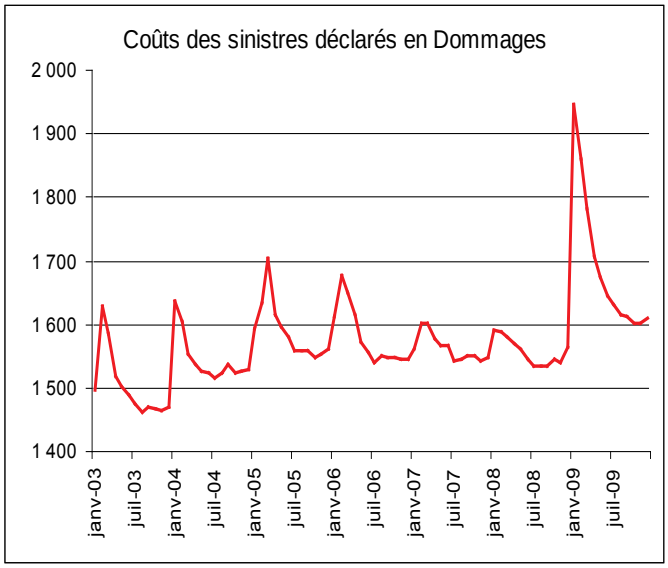


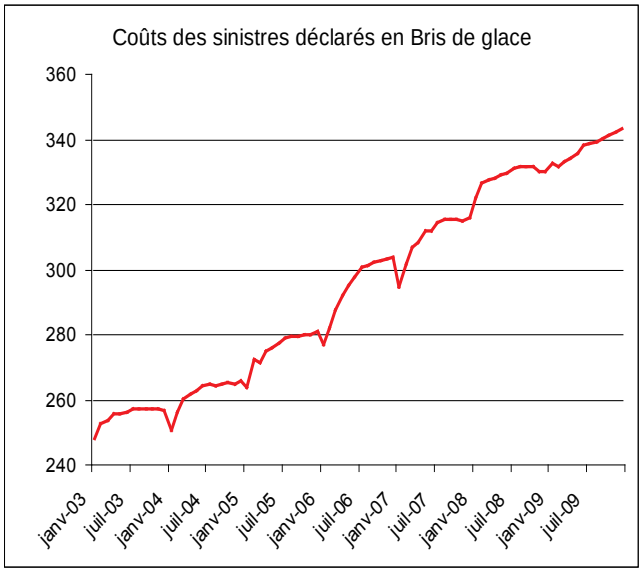


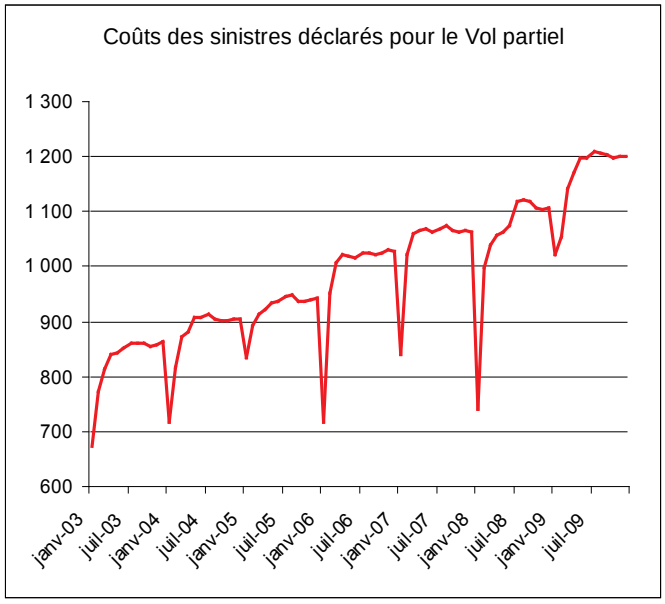


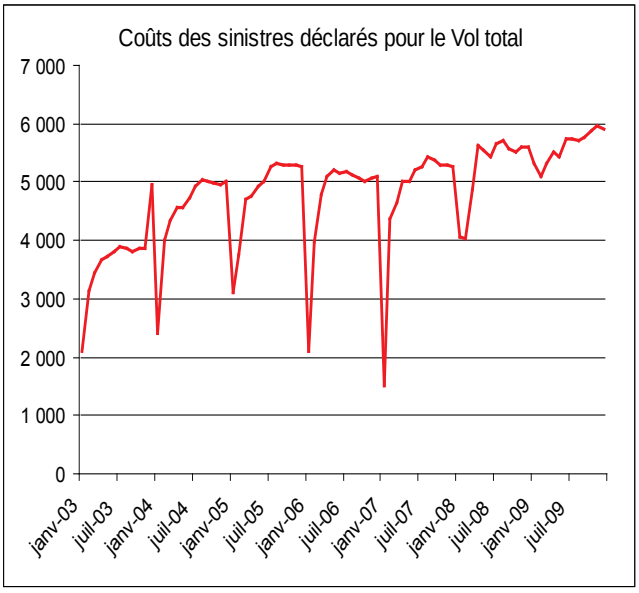


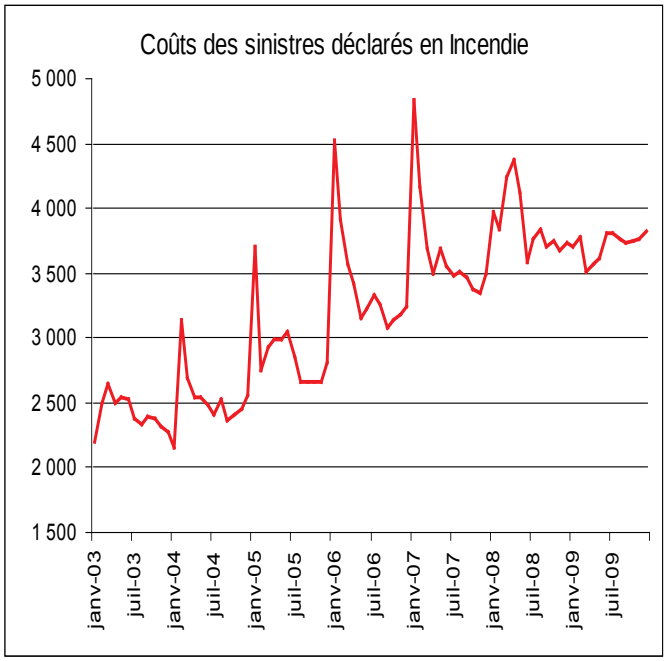


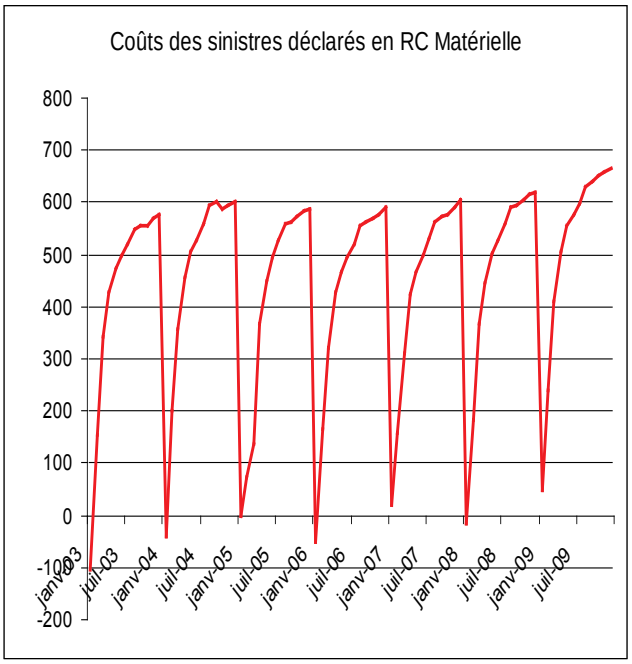

Coûts des sinistres déclarés en Dommages
2 000
1 900
1 800
1 700
1 600
1 500
1 400
janv-03
juil-03
janv-04
juil-04
janv-05
juil-05
janv-06
juil-06
janv-07
juil-07
janv-08
juil-08
janv-09
juil-09


Coûts des sinistres déclarés en Bris de glace
360
340
320
300
280
260
240
janv-03
juil-03
janv-04
juil-04
janv-05
juil-05
janv-06
juil-06
janv-07
juil-07
janv-08
juil-08
janv-09
juil-09


Coûts des sinistres déclarés pour le Vol partiel
1 300
1 200
1 100
1 000
900
800
700
600
janv-03
juil-03
janv-04
juil-04
janv-05
juil-05
janv-06
juil-06
janv-07
juil-07
janv-08
juil-08
janv-09
juil-09


Coûts des sinistres déclarés pour le Vol total
7 000
6 000
5 000
4 000
3 000
2 000
1 000
0
janv-03
juil-03
janv-04
juil-04
janv-05
juil-05
janv-06
juil-06
janv-07
juil-07
janv-08
juil-08
janv-09
juil-09


Coûts des sinistres déclarés en Incendie
5 000
4 500
4 000
3 500
3 000
2 500
2 000
1 500
janv-03
juil-03
janv-04
juil-04
janv-05
juil-05
janv-06
juil-06
janv-07
juil-07
janv-08
juil-08
janv-09
juil-09


Coûts des sinistres déclarés en RC Matérielle
800
700
600
500
400
300
200
100
0
-100
-200
janv-03
juil-03
janv-04
juil-04
janv-05
juil-05
janv-06
juil-06
janv-07
juil-07
janv-08
juil-08
janv-09
juil-09

ANNEXE 2 : Mensualisation des séries

Certaines des variables à notre disposition étaient renseignées sur une fréquence annuelle, ou trimestrielle. Pour uniformiser la base de données, il a fallu retravailler ces séries et les rendre mensuelles. La méthode d'interpolation utilisée a été celle incrémentée sous SAS avec la procédure EXPAND.

Voici le code utilisé pour passer de séries trimestrielles à des séries annuelles :

Code SAS 1

```
proc expand data=stage.immatr3 out=stage.immat3 from=qtr to=month;
convert tous_vehicules /observed=total method=step;
id date;
run;
```

Avec l'option observed = total, on précise que les valeurs originelles représentent le total sur le trimestre ; et avec l'option method = step, on obtiendra pour chaque mois la valeur moyenne sur le trimestre, en tenant compte du nombre de jours des différents mois.

Les données portefeuille ont subi le même traitement avec la procédure EXPAND, pour passer de séries annuelles, à des séries mensuelles. Toutefois, nous n'avons fait que répliquer le même nombre d'années police pour tous les mois d'une même année. En effet, l'information dont on dispose en fin d'année tient déjà compte du fait que les assurés ont pu changer de situation au cours de l'année. On a déjà la durée d'exposition réelle de l'assuré ; diviser par 12 serait déformer la durée d'exposition réelle.

Le fichier sur les vitesses moyennes de jour et de nuit sur les différents axes a dû être retraité. En effet, nous ne disposions que de la moyenne sur chaque quadrimestre. Cette fois, nous n'avons pas utilisé la procédure EXPAND, mais nous avons opté pour la démultiplication de la même valeur pour chacun des mois d'un même quadrimestre. L'idée étant de rester le plus fidèle possible à la vraie vitesse selon les périodes de l'année.

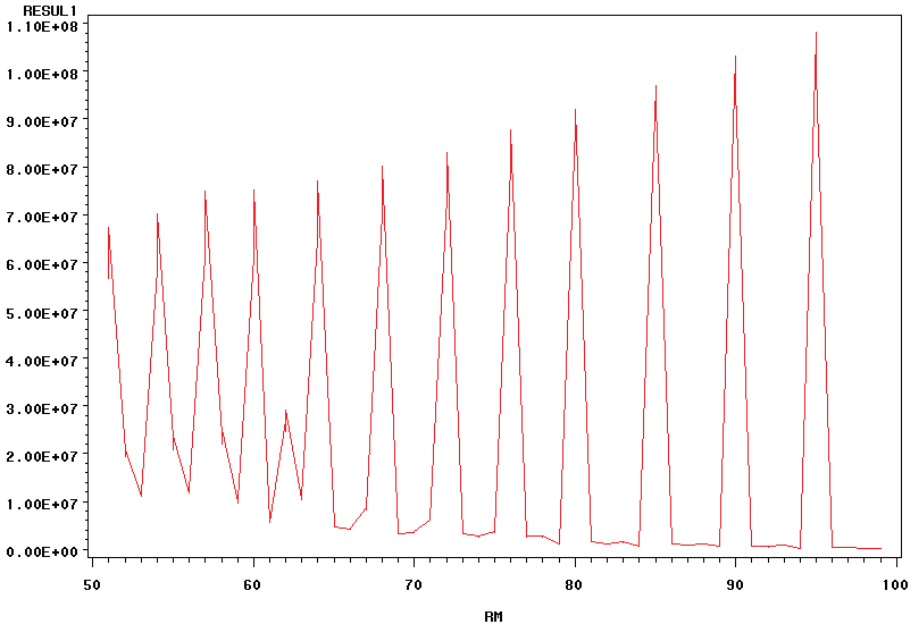
Ce traitement nous a permis de fusionner nos différentes familles de données et d'obtenir la base finale, qui n'est constituée que de données mensuelles, de janvier 2003 à décembre 2009.

ANNEXE 3 : Regroupement en classes des variables portefeuille

Cas du R/M

Nous avons tenu à conserver les classes suivantes : Bonus 50, Bonus 100, Bonus supérieur à 100. En effet, les Bonus 50 représentent 60% du portefeuille d'assurés auto AXA.

Nous avons ensuite découpé la classe des RM 51-99 en des classes plus fines, selon la durée d'exposition. Pour cela, nous avons visualisé la répartition des années police selon le RM, via une procédure GPLOT.



Pour garder approximativement 5% de l'effectif total dans chaque classe, la décomposition qui en ressortait était la suivante : une classe [51 ; 60] et une deuxième classe [61, 99]. Nous avons ensuite redécoupé la classe 66-99 en deux classes : [61-74] et [75-99].

Conversion des classes APSAD en classes SRA

Nous avons d'abord réduit le nombre de classes SRA dont nous disposions pour passer de 14 classes à 7 classes réparties comme suit :

SRA 1 = somme des classes A à D

SRA 2 = somme des classes E à H

SRA 3 = somme des classes I et J

SRA 4 = somme des classes K et L

SRA 5 = somme des classes M et N

SRA 6 = somme des classes O à V

SRA 7 = somme des classes W à Z5

Ensuite, nous avons comparé le fichier RT de 2004 et celui de 2008. En 2004, le système de classes de prix était celui de l'APSAD ; tandis qu'en 2008, on était déjà passé au système SRA. Après avoir trié nos données selon le code GTA du véhicule, nous avons fusionné nos tables, en gardant la correspondance par code GTA. Pour finir, une procédure FREQ sur les croisements SRA/APSAD nous a permis d'obtenir la correspondance suivante :

Classes APSAD Y à A (0 à 6 et 7 à 9) → SRA 1

Classes APSAD A (10 et +) à B → SRA 2

Classes APSAD C → SRA 3

Classes APSAD D à E → SRA 4

Classes APSAD F → SRA 5

Classes APSAD G, H, J, K, L, M, N, P, R → SRA 6

Classes APSAD S à T → SRA 7

ANNEXE 4 : Etude des résidus

Dans tous les modèles cités précédemment, une attention particulière a été portée aux résidus. En un point d'observation i, l'écart entre Y observé et Y estimé par le modèle est le résidu au point i : $e_i = Y_i - \hat{Y}_i$. Ces résidus e_i sont vus comme les erreurs observées des vraies erreurs inconnues ε_i telles que $\varepsilon_i = Y_i - E(Y_i)$.

Les hypothèses faites sur les ε_i pour élaborer les tests statistiques se résument ainsi « les erreurs doivent être indépendantes et identiquement distribuées ». Dans les méthodes d'estimation utilisant la méthode des moindres carrés, il faut en plus que les résidus suivent une loi normale. En effet, les résidus contiennent d'une part un aléa d'espérance nulle et de variance σ^2, et d'autre part une information concernant l'inadéquation du modèle aux données (c'est-à-dire l'écart entre le modèle postulé et le modèle correct inconnu). Ce que l'on veut c'est que l'importance de cette deuxième partie soit moindre que celle due à l'aléa.

Si le modèle est approprié aux données, les résidus observés e_i doivent refléter les propriétés des vraies erreurs inconnues ε_i. On doit donc s'assurer de la conformité de ces hypothèses.

1. Tests de bruit blanc

1. a. Fonction d'autocorrélation empirique (ACF)

Nous rappelons que la fonction d'autocorrélation d'une suite de variables aléatoires indépendantes et identiquement distribuées de variance finie, vérifie $\rho(h) = 0, \forall h > 0$. Son estimée, l'ACF $\hat{\rho}$ doit par conséquent être proche de 0. On a le résultat suivant :

Soit $Y_1, \ldots, Y_n$ v.a. i.i.d. de variance finie. Alors $(\hat{\rho}_Y^{(n)}(h), h > 0)$ est approximativement une suite de v.a. i.i.d. $N(0, 1/n)$, pour n grand.

Par conséquent approximativement 95% des $\hat{\rho}^{(n)}$ devraient se situer entre les bornes $\pm \frac{q_{(0,975)}}{\sqrt{n}} = \pm \frac{1,96}{\sqrt{n}}$ pour satisfaire l'hypothèse d'indépendance.

En pratique, si l'on calcule $\hat{\rho}^{(n)}(h)$, avec h = 1,…, 40, et que l'on trouve plus de 2 (= 40 * 0,05) ou 3 valeurs en dehors de ces bornes (ou qu'une valeur soit vraiment très loin des bornes), alors on rejettera l'hypothèse d'avoir une suite i.i.d.

1. b. Test du Porte-manteau

On considère la statistique associée à l'ACF $\hat{\rho}^{(n)}(.)$ définie par :

$$Q_n = n \sum_{h=1}^{k} (\hat{\rho}^{(n)}(h))^2$$

On remarque, en utilisant le modèle $(*)$, que si $Y_1,\ldots, Y_n$ est une suite de v.a. i.i.d. de variance finie, alors la loi de Q_n est, pour n grand, approximativement une loi du khi-deux à k degrés de liberté. Une grande valeur de Q_n nous indique donc que les autocorrélations des données sont trop grandes pour que les données soient celles d'une suite i.i.d. d'où le test au niveau de région critique (i.e. rejet de l'hypothèse i.i.d.) $Q_n > \chi_k^2 (1-\alpha)$; avec $\chi_k^2 (1-\alpha)$ = quantile d'ordre $(1-\alpha)$ du Khi-deux d'ordre k.

2. Tests de normalité des résidus

Nous utilisons la technique graphique du Q-Q plot et le test de Shapiro-Wilk. Le seuil limite retenu est 5%.

2. a. Le Q-Q plot :

Le Q-Q plot, *quantile-quantile plot* est une technique graphique qui permet de comparer les distributions de deux ensembles de données. Dans notre cas, nous voulons tester l'ajustement de nos résidus à une loi normale N (0, 1).

Soit $x_{(i)}$ tel que $x_{(1)} < x_{(2)} < \ldots.. < x_{(n)}$. On associe à chaque valeur $x_{(i)}$, le i/(n+1)-quantile d'une loi normale centrée réduite noté $x_{(i)}$*.

Si les données sont compatibles avec une loi normale, les points ($x_{(i)}$, $x_{(i)}$*) forment une droite dite *droite de Henry* ; ils sont alignés sur la diagonale principale.

2. b. Le test de Shapiro – Wilk :

Il est basé sur la statistique W. Il teste si la réalisation $x_1, x_2, \ldots, x_n$ de la suite de variables aléatoires $X_1, \ldots, X_n$ est normalement distribuée. Il est particulièrement puissant pour les petits échantillons de taille ≤ 50. La statistique W est la suivante :

$$W = \frac{\left[\sum_{i=1}^{[n/2]} a_i (x_{(n-i+1)} - x_{(i)})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

où $x_{(i)}$ est la $i^{ème}$ observation de la série des observations ordonnées ;

[n/2] est la partie entière du rapport n/2 ;

les termes a_i sont des constantes générées à partir de la moyenne et de la matrice de covariance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont disponibles dans des tables spécifiques.

On rejette l'hypothèse de normalité si $W < W_{critique}$. Les seuils $W_{critique}$ pour différents risques α et effectifs n sont disponibles dans la table de Shapiro-Wilk.

3. Test de Durbin-Watson

Ce test est utilisé pour détecté la présence d'autocorrélation de type autorégressif dans des résidus issus d'une régression. Il part de l'hypothèse que les résidus sont stationnaires et distribués selon une loi normale centrée. Il teste les hypothèses

H_0 : les erreurs sont non corrélées

H_1 : les erreurs sont un AR(1), c'est-à-dire corrélées à l'ordre 1

La statistique de Durbin-Watson est la suivante :

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$, avec e_t le résidu

Si les erreurs sont un bruit blanc, d sera très proche de 2 et si elles sont fortement autocorrélées, d sera très éloignée de 2.

La statistique de Durbin-Watson de base ne testait que l'autocorrélation au retard d'ordre 1. Elle a été généralisée pour tester l'autocorrélation à des retards plus élevés (H_1 : les erreurs sont un AR(k)).

La statistique de Durbin - Watson généralisée est la suivante :

$$d_k = \frac{\sum_{t=k+1}^{n}(e_t - e_{t-k})^2}{\sum_{t=1}^{n} e_t^2}$$

Pour chaque ordre, et ce jusqu'au $k^{ième}$ ordre, on calcule la probabilité marginale avec un système de matrice. Si la probabilité marginale basée sur la statistique d_k est plus petite que le niveau de significativité α, cela indique une autocorrélation positive au retard d'ordre k.

Par contre, on conclue à la présence d'une autocorrélation négative au retard d'ordre k si la probabilité marginale est plus grande que $1 - \alpha$.

ANNEXE 5 : Test de Dickey – Fuller

Le test de Dickey-Fuller permet de tester la présence de racines unités d'un polynôme autorégressif, afin de savoir si on doit différencier ou non la série de données. On suppose que le processus X est centré (si X n'est pas centré, on peut toujours se ramener à ce cas en posant Z = X – E(X)).

À l'ordre 1 :

Soit $X_1,\ldots, X_n$ observations d'un modèle AR(1) : $X_t + \phi_1 X_{t-1} = \varepsilon_t$ (1)

avec (ε_t) b.b. de variance σ^2. On veut construire un test :

$H_0 : |\phi_1| = 1$

$H_1 : |\phi_1| < 1$

Si $|\phi_1| < 1$, pour n assez grand, l'EMV $\hat{\phi}_1$ de ϕ_1 suit approximativement une loi

$N(-\phi_1, (1-\phi_1^2)/n)$.

Si $|\phi_1| \cong 1$, cette approximation normale n'est pas possible, même dans le cas asymptotique, ce qui exclut son utilité pour construire le test d'hypothèse nulle H_0.

Construction du test d'hypothèse nulle H_0 par Dickey et Fuller:

Le modèle (1) peut s'écrire sous la forme $\nabla X_t = X_t - X_{t-1} = \phi_1^* X_{t-1} + \varepsilon_t$ (2),

avec $\phi_1^* := -1 - \phi_1$

Soit $\hat{\phi}_1^*$ l'EMC de ϕ_1^* obtenu par régression de ∇X_t sur X_{t-1}.

L'erreur estimée de $\hat{\phi}_1^*$, notée $\hat{SE}(\hat{\phi}_1^*)$, est alors :

$$\hat{SE}(\hat{\phi}_1^*) = \frac{S}{\sqrt{\sum_{t=2}^{n}(X_{t-1} - \bar{X}_{n-1})^2}} \quad \text{avec } S^2 = \frac{\sum_{t=2}^{n}(\nabla X_t - \hat{\Phi}_1^* X_{t-1})^2}{n-3}.$$

On considère alors le t-ratio $\hat{\tau}_1 = \frac{\hat{\Phi}_1^*}{\hat{SE}(\hat{\Phi}_1^*)}$.

Dickey et Fuller ont calculé la distribution limite de $\hat{\tau}_1$ quand $n \to \infty$ sous l'hypothèse $\Phi_1^* = 0$ (i.e. existence d'une racine unité), ce qui a permis la construction d'un test d'hypothèse H_0 : $|\Phi_1| = 1$. Les 0.01, 0.05 et 0.1 quantiles de la distribution limite de $\hat{\tau}_1$ sont respectivement -3,43, -2,86 et -2,57.

Le test de D.F. rejette donc H_0 (i.e. existence de racine unité) au niveau 0.05 si $\hat{\tau}_1 <$ 2,86. On remarque que cette valeur (-2,86) est beaucoup plus petite que celle correspondant à une approximation normale (-1,45); on rejettera donc moins facilement l'existence d'une racine unité en utilisant la loi asymptotique correcte.

<u>À l'ordre p :</u>

La procédure précédente peut être généralisée à l'ordre p.

On part du modèle $X_t + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} = \varepsilon_t$ (3), avec (ε_t) b.b. de variance σ^2.

On l'écrit alors sous la forme :

$$\nabla X_t = \phi_1^* X_{t-1} + \phi_2^* \nabla X_{t-1} + \ldots + \phi_p^* \nabla X_{t-p+1} + \varepsilon_t \quad (4)$$

avec $\phi_1^* := -1 - \sum_{i=1}^{p} \phi_i$ et $\phi_j^* := \sum_{i=j}^{p} \phi_i$, $2 \le j \le p$.

Si le polynôme autorégressif a une racine unité en 1, alors $0 = \phi(1)$, et donc $\phi_1^* = 0$, d'où l'on déduit que le processus différencié ∇X_t est un AR(p – 1).

Par conséquent, tester l'hypothèse d'une racine unité du polynôme autorégressif en 1 revient à tester $\phi_1^* = 0$. Comme pour le cas de l'AR(1), on estime ϕ_1^* par l'EMC obtenu par régression de ∇Xt sur X_{t-1}, ∇X_{t-1}, …, ∇X_{t-p}.

Pour n assez grand, le t-ratio $\hat{\tau}_p := \dfrac{\hat{\Phi}_1^*}{\hat{SE}(\hat{\Phi}_1^*)}$ a la même distribution asymptotique que la statistique de test $\hat{\tau}_1$. On peut donc construire le même test (avec les mêmes valeurs de quantiles) que dans le cas de l'AR(1) en utilisant cette fois la statistique $\hat{\tau}_p$.

On a supposé dans la théorie ci-dessus les aléas ε_t i.i.d et notamment indépendants. Cette hypothèse très forte est le plus souvent irréaliste. Autorisant l'aléa lui-même à suivre un modèle autorégressif, Dickey et Fuller ont déterminé les modèles transformés correspondants et calculé les nouvelles tables permettant de faire les tests précédents. C'est ce qui nous a fourni les tests et tables de Dickey-Fuller augmentés (ADF). Ils ne seront pas détaillés ici.

ANNEXE 6 : Théorie de la régression robuste avec les MM-estimateurs

La MM-estimation, développé par Victor Yohai (1987), est une combinaison de l'estimateur à fort point de rupture (High breakdown estimator LTS ou S-estimateur) et du M-estimateur. Son but est de conserver la robustesse d'un estimateur à haut point de rupture et l'efficacité d'un M-estimateur. La caractéristique principale d'une méthode robuste est son "point de rupture" ("breakdown point" en anglais). Le point de rupture est le pourcentage de mesure aberrante qui met en défaut l'algorithme. À titre d'exemple, l'algorithme des moindres carrés a un point de rupture de 0% car il suffit d'une seule mesure aberrante pour obtenir une estimation fausse des paramètres.

La MM-estimation est basée sur la minimisation de la somme des résidus centrés et transformés. La fonction de transformation est choisie sans contrainte. Mais en général, c'est une fonction qui croit lentement, comme la fonction bicarrée de Tukey.
Dans la méthode des MCO, la fonction de transformation est la fonction carrée.

L'estimation se fait en trois étapes :

1. les valeurs initiales de l'estimateur sont celles de l'estimateur à fort point de rupture
2. les valeurs initiales de l'estimateur d'échelle proviennent de la résolution numérique de l'équation : $\frac{1}{n-p-1}\sum_{i=1}^{n}\chi\left(\frac{\hat{\varepsilon}_i}{\hat{\sigma}_0}\right)=\alpha$

   où $\alpha = E\varphi(\chi(\varepsilon)) = \int_{-\infty}^{\infty}\chi(\varepsilon)\varphi(\varepsilon)d\varepsilon$ est la valeur attendue des résidus χ-transformés normalement distribués.
3. Avec ces valeurs initiales, un M-estimateur local est calculé. Parce que les paramètres estimés dépendent des résidus et que les résidus eux-mêmes dépendent des paramètres estimés, un algorithme itératif appelé IRLS est utilisé pour l'estimation.

ANNEXE 7 : Théorie des modèles espace-état et du filtre de Kalman

L'étude de systèmes physiques émettant au cours du temps des signaux déterminés par des états internes non observés, a conduit à développer en traitement du signal les modèles dits espace-état. Nous avons utilisé ces modèles pour modéliser des séries temporelles. L'estimation de tels modèles se fait en deux temps : d'abord l'estimation des variables cachées (avec le filtre de Kalman), puis celle des paramètres (avec l'algorithme EM).

Le modèle peut être décrit avec les notations suivantes :

Y_t est la variable de mesure

Z_t est la variable d'état à la date t

ε_t est le vecteur des innovations à la date t

η_t est le vecteur des erreurs de mesure à la date t

A_t est la matrice de transition

C_t est la matrice de mesure

$X_{1,t}$, $X_{2,t}$ sont des variables exogènes prédéterminées

C_tZ_t est le signal à la date t.

Soit un processus multidimensionnel Y_t ; on appelle modèle espace-état de ce processus le système (I) décrit par les équations matricielles (1) et (2) :

$$\begin{cases} (1): Z_{t+1} = A_t Z_t + B_t X_{1,t} + \varepsilon_t \\ (2): Y_t = C_t Z_t + D_t X_{2,t} + \eta_t \end{cases} \text{ où } \begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim N\left(0, \begin{pmatrix} Q_t & S_t \\ S_t & R_t \end{pmatrix}\right) \qquad (I)$$

où les matrices A_t, C_t sont de taille $K \times K$ et $n \times K$, B_t et D_t sont des matrices déterministes de taille $K_1 \times K$ et $K_2 \times K$ et Z_0 est un vecteur aléatoire de loi N(m,P) indépendant du bruit blanc normal.

Les variables d'état et de mesure peuvent s'écrire en fonction de la variable d'état initiale, du passé des erreurs de mesure et des innovations ainsi que des variables exogènes. C'est la forme dite **développée** du système qui est utilisée lorsqu'on s'intéresse à l'estimateur des moindres carrés généralisés ou à l'initialisation du filtre de Kalman.

Le filtre de Kalman est une procédure récursive qui nous permet de calculer l'estimateur optimal d'un vecteur état à l'instant t, en se basant sur les informations disponibles jusqu'à cet

instant. Nous sommes ici en présence d'un problème complexe. Nous avons des paramètres inconnus et des composantes inobservables qu'il n'est pas possible d'obtenir par une quelconque inférence séparément. Le filtre de Kalman permet d'apporter une solution à ce problème du fait de ses deux particularités importantes. Grâce à la forme état-mesure linéaire et pour une valeur particulière des paramètres inconnus, le filtre de Kalman peut :

- produire une inférence concernant les composantes inobservables,
- calculer la somme de la log-vraisemblance qui est fonction des paramètres inconnus, et qui est issue de l'inférence sur les inobservables

Dès lors, le filtre de Kalman peut résoudre les deux problèmes simultanément :

(i) dans un premier temps, il permet de mener une inférence sur les paramètres inconnus par la méthode du maximum de vraisemblance. En effet, on peut grâce à lui disposer d'une expression de la vraisemblance pour le modèle mis sous la forme état-mesure linéaire.

(ii) dans un second temps : une fois les paramètres inconnus estimés, le filtre produit une inférence sur les composantes inobservables pour la solution du maximum de vraisemblance.

De façon générale, l'idée consiste à trouver les valeurs des paramètres inconnus permettant au filtre de fournir les variables inobservables permettant d'effectuer à chaque période les meilleures prévisions possibles sur les variables de mesure au sens du maximum de vraisemblance. Grâce à des hypothèses supplémentaires concernant les conditions initiales des variables d'état et conditionnellement aux paramètres à estimer, le filtre de Kalman itère sur l'échantillon en répétant quatre étapes à chaque date.

Partant de prévisions réalisées sur les variables d'état à partir des équations d'état (étape 1), on peut obtenir une prévision sur les variables de mesure grâce aux équations de mesure (étape 2). Cette prévision est préalable à la réalisation des variables de mesure observées, c'est-à-dire conditionnelle à l'information passée. Une fois l'information contemporaine acquise, l'erreur de prévision sur les mesures (étape 3) permet d'améliorer la première prévision sur les états (étape 4), qui permet ensuite d'alimenter l'itération suivante du filtre, et ainsi de suite jusqu'à la dernière observation de l'échantillon. L'algorithme consiste à estimer à chaque instant t les variables cachées (le vecteur d'état) conditionnellement aux variables observées jusqu'à la date t (le vecteur de mesure).

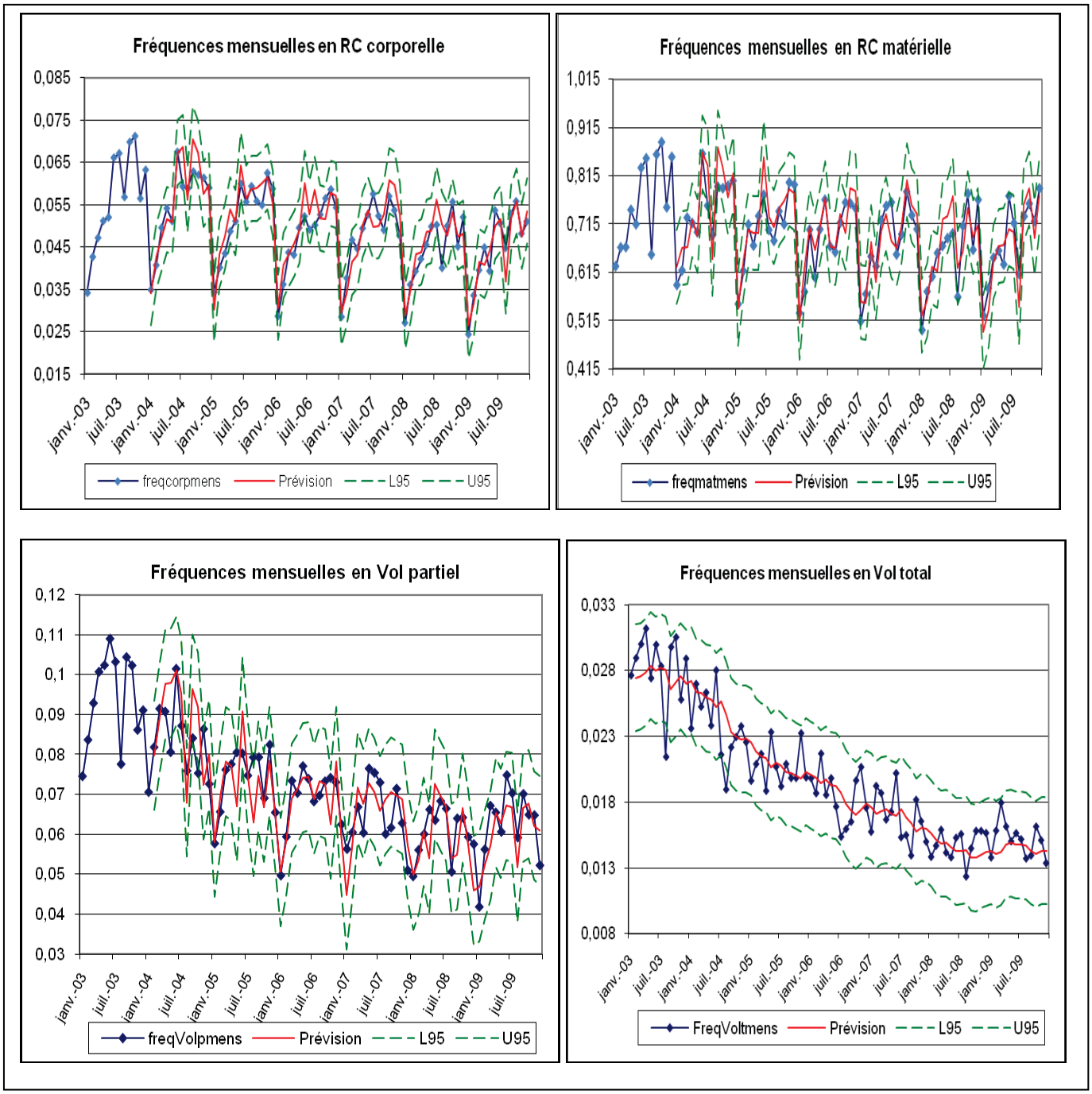
L'estimateur du vecteur d'état envisagé jusqu'ici est une estimation filtrée et se distingue des estimations lissées ou prévues :

- → le filtrage consiste à rechercher la meilleure approximation de l'état Z_t sachant les observations présentes et passées Y_0, …, Y_t.
- → la prévision consiste à rechercher la meilleure approximation de l'état Z_t sachant les observations passées Y_0, …, Y_t.
- → le lissage consiste à rechercher la meilleure approximation de l'état Z_t sachant les observations passées, présentes et futures Y_0, …, Y_t.

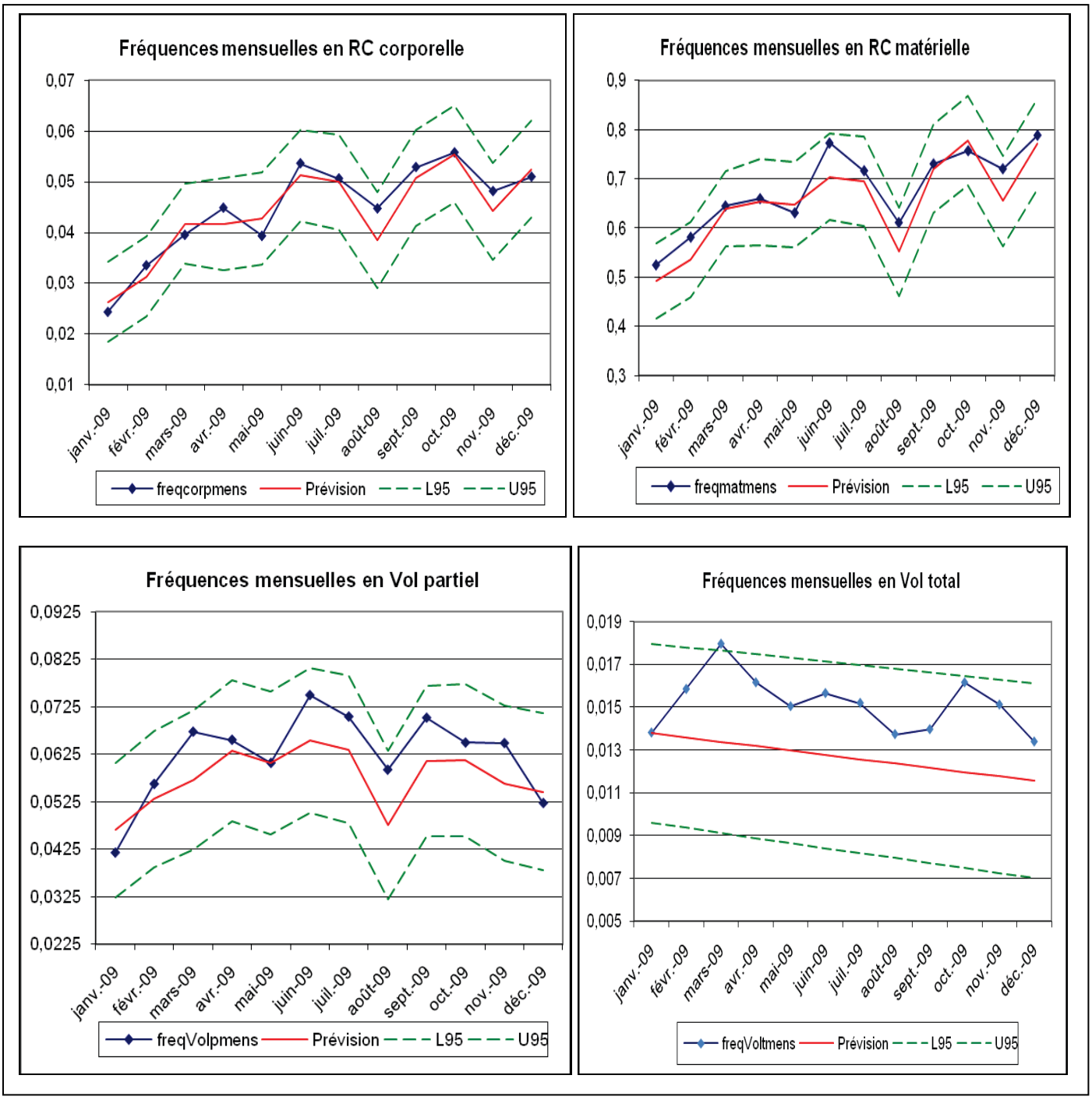
Les problèmes de prévisions et de lissage se traitent à partir de simples extensions de l'algorithme de filtrage de base.

ANNEXE 8 : Résultats de la partie modélisation ARMA

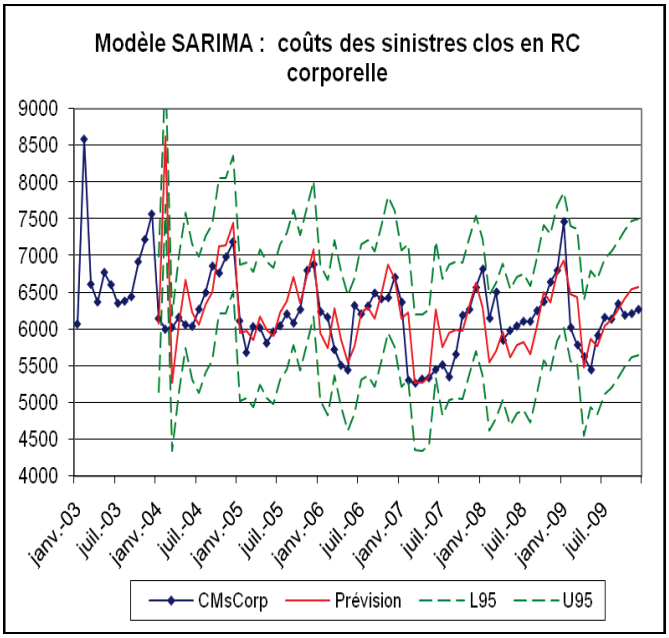
Prévisions des fréquences mensuelles avec les données complètes



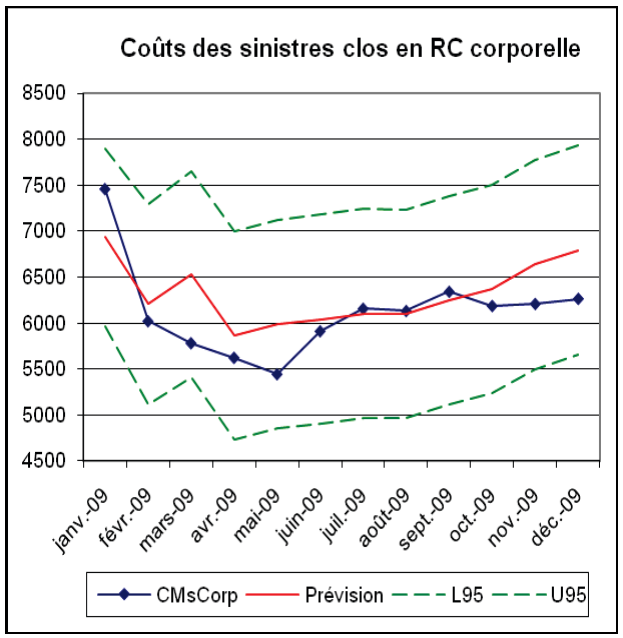
Prévisions des fréquences mensuelles pour l'année 2009, avec les données tronquées à fin 2008



Prévisions des coûts de sinistres clos

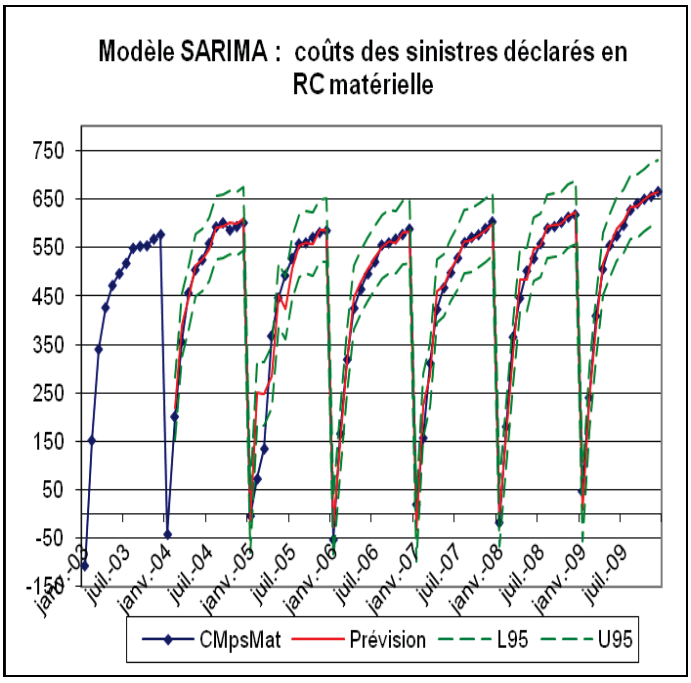


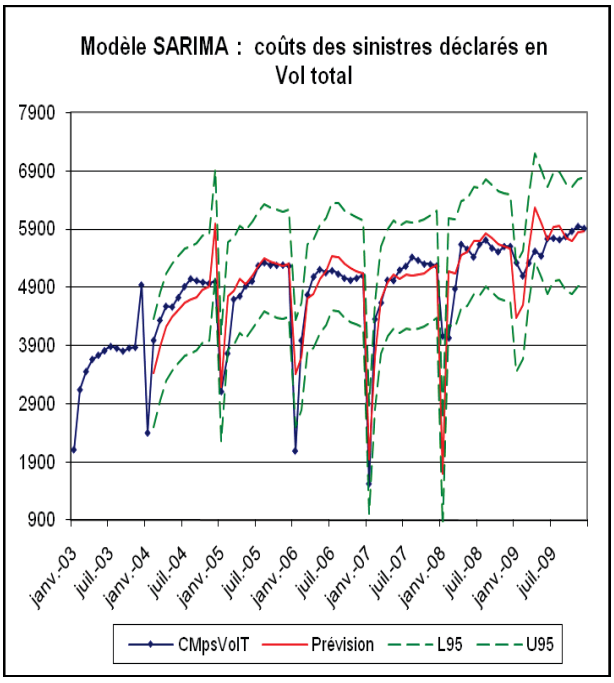
Avec les données complètes



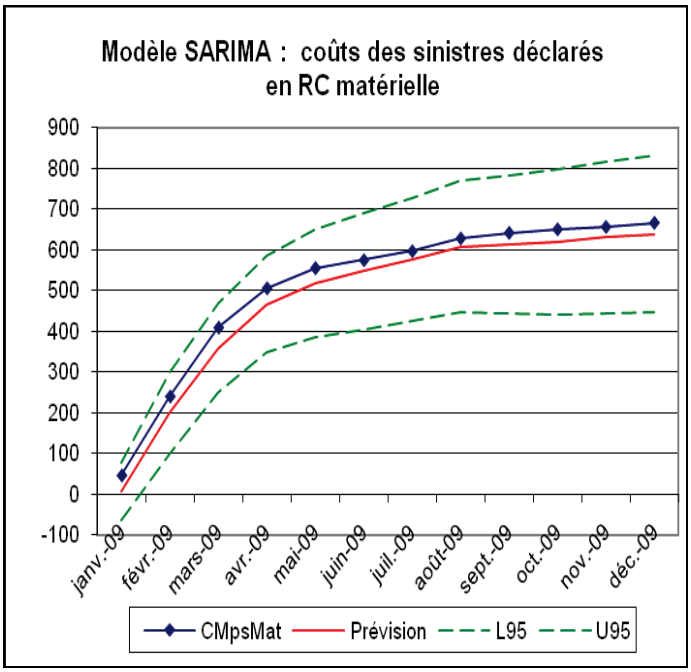
Avec les données tronquées à fin 2008

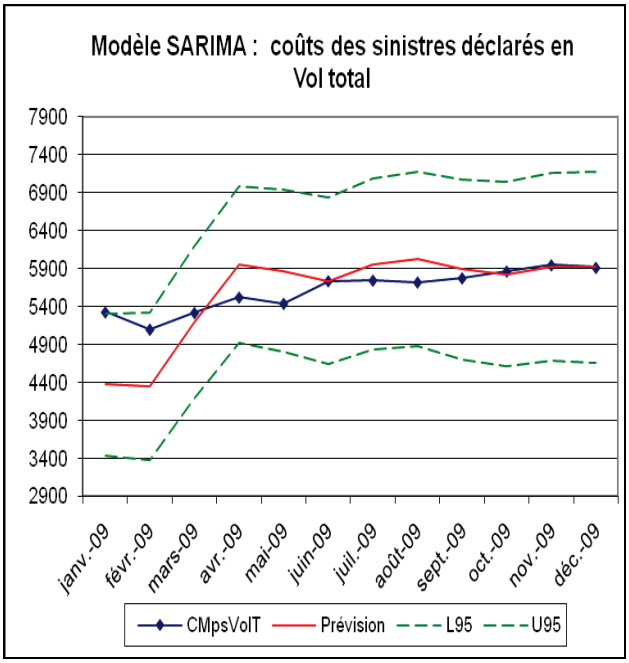
Prévision des coûts de sinistres déclarés





Avec les données complètes





Avec les données tronquées à fin 2008

Tests sur les résidus des fréquences avec la modélisation ARMA

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	1.70	6	0.9452	0.055	0.057	-0.098	0.050	0.056	-0.022
12	8.14	12	0.7738	-0.051	-0.084	0.037	0.079	-0.042	-0.233
18	21.84	18	0.2391	-0.331	0.046	0.005	0.090	-0.139	-0.100

Tests de normalité

Test	Statistique		p Value	
Shapiro-Wilk	**W**	0.991461	**Pr < W**	0.9117
Kolmogorov-Smirnov	**D**	0.06396	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.02984	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.196362	**Pr > A-Sq**	>0.2500

RC corporelle

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	2.04	6	0.9163	-0.039	0.099	-0.059	-0.088	0.029	-0.055
12	10.44	12	0.5775	-0.067	0.079	0.241	0.061	0.009	-0.161
18	19.81	18	0.3439	-0.231	0.006	-0.172	-0.010	0.107	0.083

RC matérielle

RC matérielle

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.984479	**Pr < W**	0.5198
Kolmogorov-Smirnov	**D**	0.066093	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.052642	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.309118	**Pr > A-Sq**	>0.2500

Vol partiel

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	6.92	6	0.3282	-0.095	-0.097	0.197	-0.023	0.174	0.045
12	17.16	12	0.1436	-0.154	0.045	0.201	-0.177	-0.056	-0.145

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.985836	**Pr < W**	0.6077
Kolmogorov-Smirnov	**D**	0.071793	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.069049	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.391793	**Pr > A-Sq**	>0.2500

Vol total

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	4.20	6	0.6498	0.012	-0.135	0.042	0.016	0.069	0.147
12	27.29	12	0.0070	0.015	0.106	-0.063	-0.347	0.043	0.314
18	34.25	18	0.0117	0.013	-0.157	-0.108	-0.120	0.117	-0.047

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.97265	**Pr < W**	0.0736
Kolmogorov-Smirnov	**D**	0.056245	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.03544	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.349289	**Pr > A-Sq**	>0.2500

Tests sur les résidus des coûts avec la modélisation ARMA

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	2.67	6	0.8492	-0.064	-0.015	0.148	0.088	-0.023	0.006
12	6.46	12	0.8911	-0.023	0.015	-0.082	0.085	-0.085	-0.147
18	9.42	18	0.9492	-0.161	-0.033	-0.007	-0.057	0.047	0.004

Coûts clos en RC corporelle

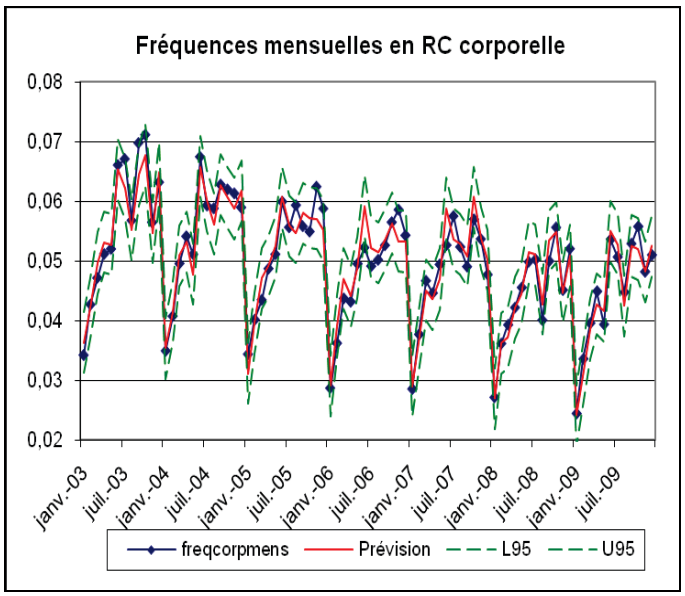
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	3.59	6	0.7316	0.056	0.135	-0.105	-0.090	-0.074	0.034
12	10.55	12	0.5682	0.101	-0.012	-0.010	0.109	0.141	-0.196

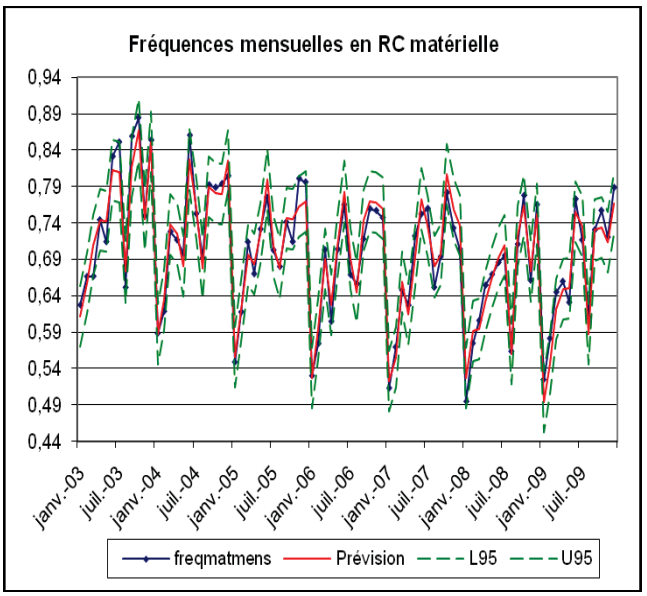
Coûts déclarés en RC matérielle

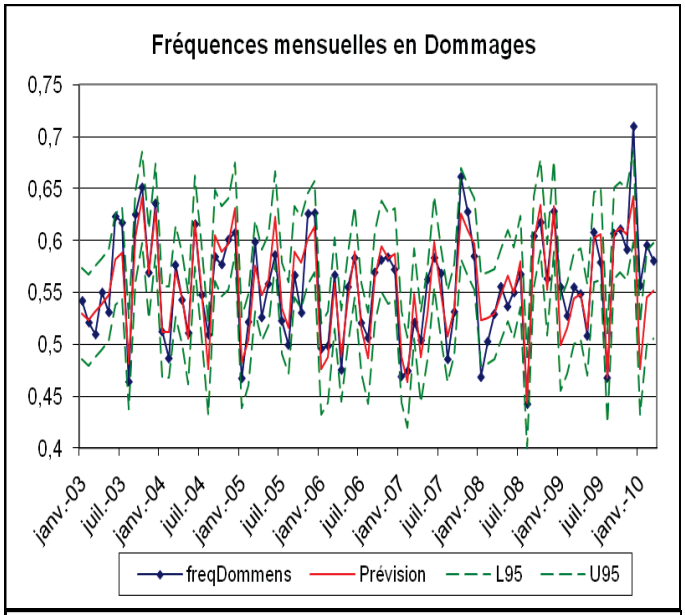
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	0.70	6	0.9945	-0.026	0.069	-0.011	0.033	-0.043	-0.026
12	1.75	12	0.9997	-0.025	-0.064	0.013	-0.052	0.050	-0.049

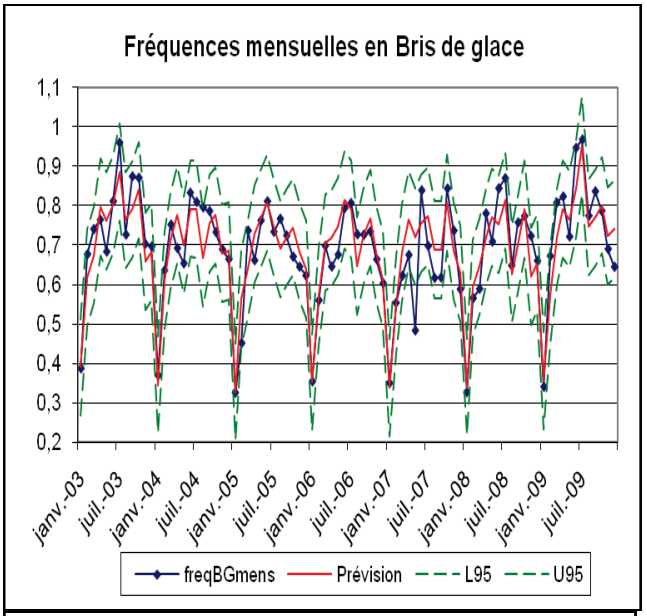
Coûts déclarés en Vol total

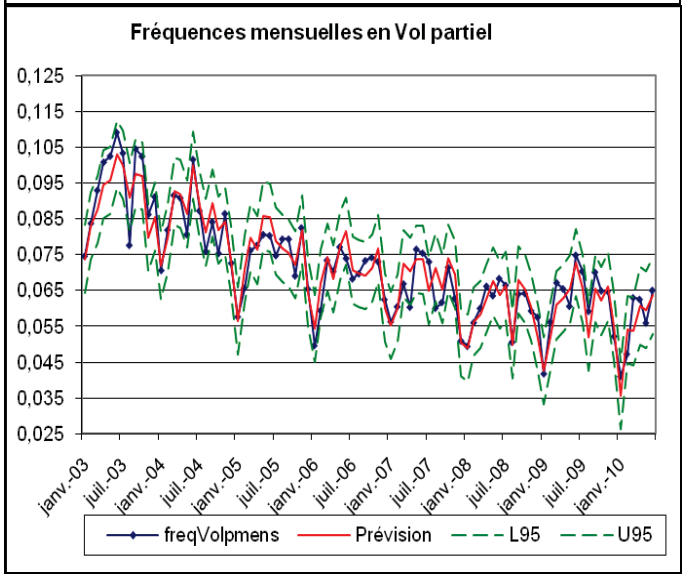
ANNEXE 9 : Résultats de la partie régression avec erreurs ARMA

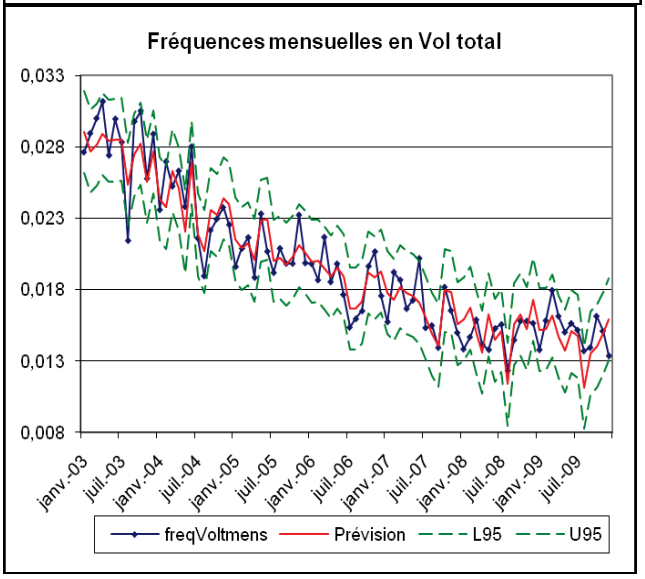












Tests sur les résidus des modèles de régression avec erreurs ARMA pour les fréquences

RC corporelle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	6.44	6	0.3758	0.130	0.060	0.011	-0.018	-0.200	-0.102
12	16.94	12	0.1519	-0.240	-0.079	0.028	0.030	0.054	0.201
18	29.80	18	0.0394	-0.111	-0.033	0.302	0.060	-0.119	0.022

Tests de normalité				
Test	Statistique		p Value	
Shapiro-Wilk	W	0.986629	Pr < W	0.5400
Kolmogorov-Smirnov	D	0.058294	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.029664	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.25126	Pr > A-Sq	>0.2500

RC matérielle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	5.54	6	0.4767	0.059	0.222	0.003	-0.071	-0.038	-0.056
12	13.85	12	0.3106	-0.147	0.077	0.164	0.061	0.164	-0.032
18	16.79	18	0.5376	-0.056	-0.029	-0.096	0.029	0.056	0.103

Tests de normalité				
Test	Statistique		p Value	
Shapiro-Wilk	W	0.984477	Pr < W	0.4100
Kolmogorov-Smirnov	D	0.055594	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.056366	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.358884	Pr > A-Sq	>0.2500

Dommages

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	0.82	6	0.9915	0.004	-0.001	-0.011	-0.094	-0.005	0.010
12	8.66	12	0.7316	-0.143	-0.101	0.092	-0.040	0.112	-0.165
18	11.63	18	0.8657	-0.081	0.038	0.036	0.110	-0.011	0.080

Dommages

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.978664	**Pr < W**	0.1771
Kolmogorov-Smirnov	**D**	0.068845	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.070181	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.490761	**Pr > A-Sq**	0.2222

Bris de glace

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	6.99	6	0.3220	0.084	0.154	0.049	0.079	0.196	0.009
12	18.76	12	0.0946	0.202	-0.008	-0.012	0.102	-0.008	0.262
18	24.39	18	0.1425	-0.122	-0.043	-0.106	-0.105	0.095	-0.074

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.972595	**Pr < W**	0.0699
Kolmogorov-Smirnov	**D**	0.044388	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.025954	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.268767	**Pr > A-Sq**	>0.2500

Vol partiel

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	3.71	6	0.7164	-0.081	-0.187	0.060	-0.029	0.057	-0.088
12	5.70	12	0.9306	-0.053	0.057	-0.001	-0.138	-0.041	-0.019

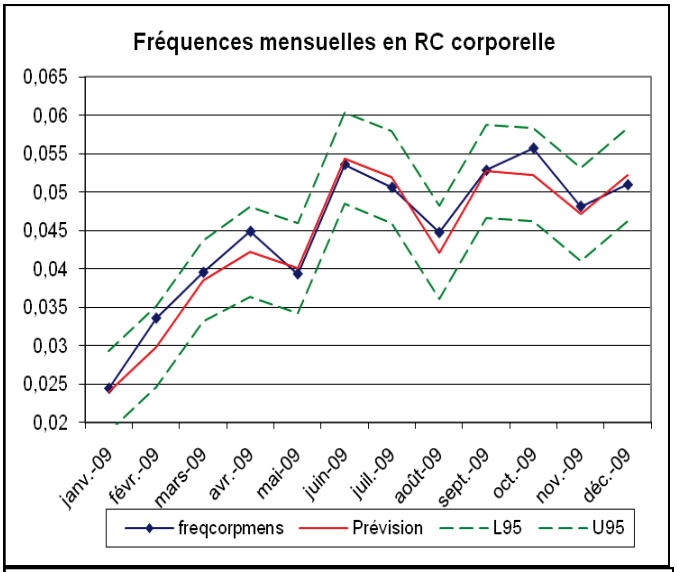
Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.980356	**Pr < W**	0.4438
Kolmogorov-Smirnov	**D**	0.084458	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.060367	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.361791	**Pr > A-Sq**	>0.2500

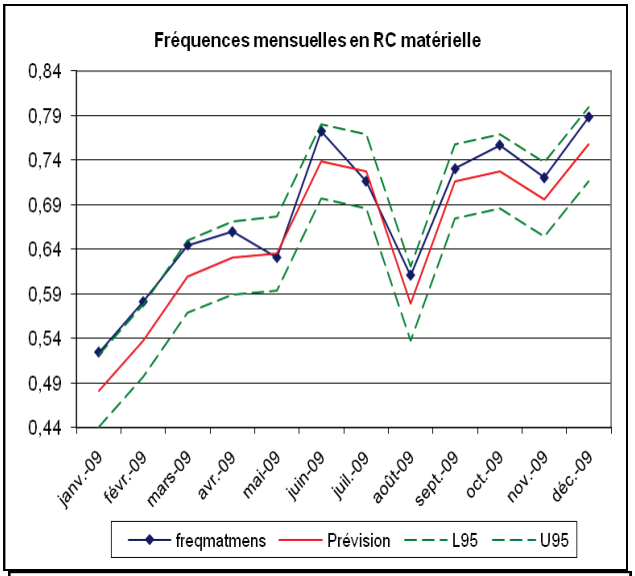
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	3.40	6	0.7572	0.064	-0.035	0.015	0.178	-0.094	-0.071
12	15.40	12	0.2201	-0.171	0.131	-0.073	-0.246	-0.085	0.206

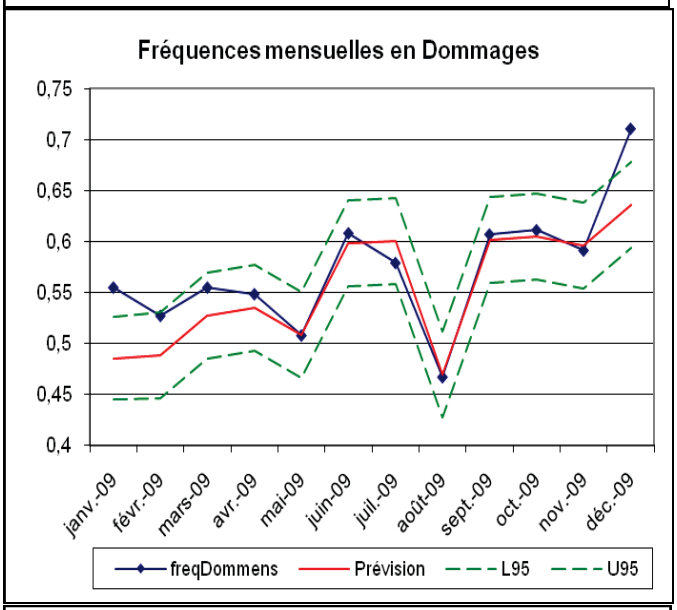
Tests de normalité				
Test	Statistique		p Value	
Shapiro-Wilk	W	0.971954	Pr < W	0.1817
Kolmogorov-Smirnov	D	0.111943	Pr > D	0.0610
Cramer-von Mises	W-Sq	0.09788	Pr > W-Sq	0.1204
Anderson-Darling	A-Sq	0.575467	Pr > A-Sq	0.1343

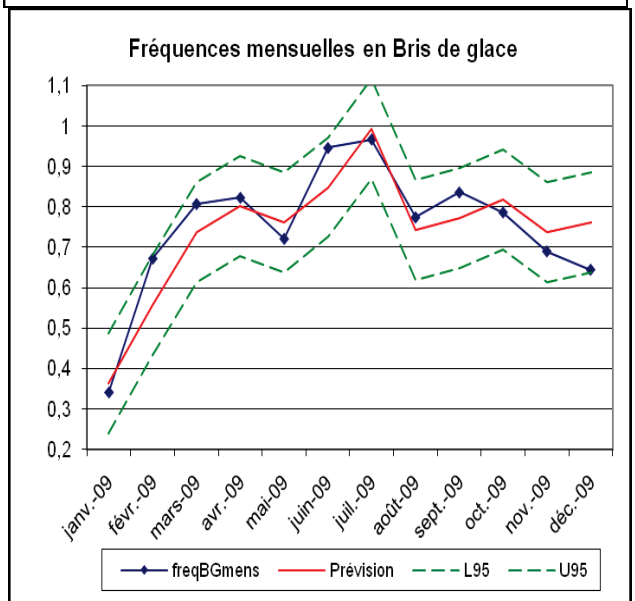
Vol total

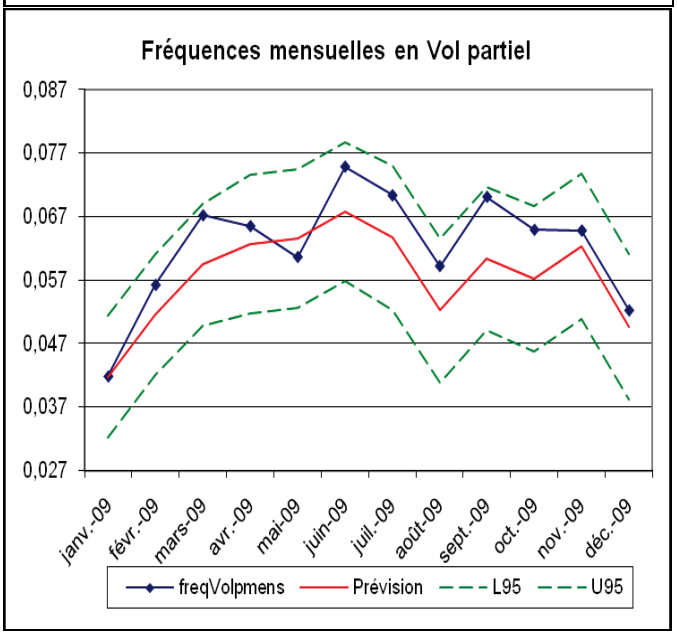
Prévisions des fréquences pour l'année 2009, avec les données tronquées à fin 2008

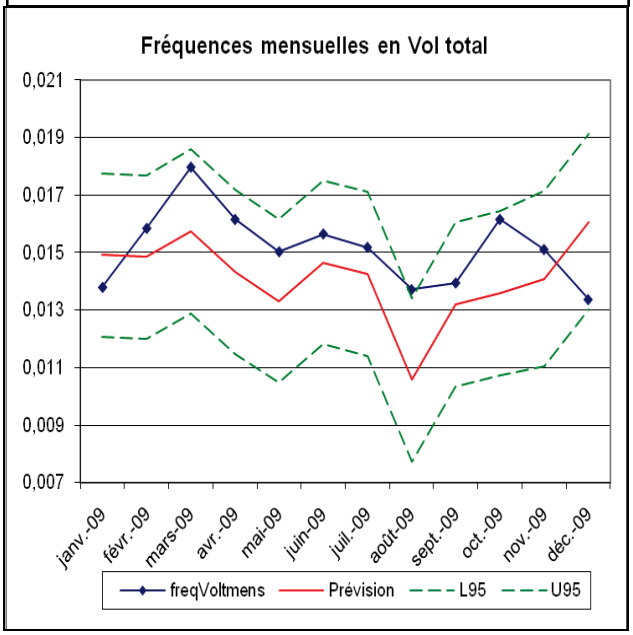












ANNEXE 10 : Résultats de la partie modélisation UCM

Tests sur les résidus des modèles UCM saturés pour les séries de fréquences

RC corporelle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	7.60	6	0.2692	0.031	-0.151	0.158	-0.152	-0.156	-0.039
12	13.65	12	0.3236	-0.173	-0.127	0.052	0.071	0.132	-0.026
18	20.45	18	0.3081	-0.116	0.137	0.051	0.011	-0.083	-0.173

Tests de normalité				
Test	Statistique		p Value	
Shapiro-Wilk	W	0.983502	Pr < W	0.4668
Kolmogorov-Smirnov	D	0.085631	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.097941	Pr > W-Sq	0.1207
Anderson-Darling	A-Sq	0.52972	Pr > A-Sq	0.1781

RC matérielle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	11.82	6	0.0661	-0.234	0.170	-0.196	-0.024	-0.048	-0.185
12	20.50	12	0.0581	-0.001	-0.077	0.227	0.060	-0.031	0.205

Tests de normalité				
Test	Statistique		p Value	
Shapiro-Wilk	W	0.985252	Pr < W	0.5931
Kolmogorov-Smirnov	D	0.097477	Pr > D	0.1003
Cramer-von Mises	W-Sq	0.068467	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.354622	Pr > A-Sq	>0.2500

Dommages

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	2.22	6	0.8980	0.018	-0.112	-0.019	-0.084	0.025	0.094
12	14.21	12	0.2877	-0.198	-0.097	0.279	0.021	0.102	0.099

Dommages

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.983764	**Pr < W**	0.5107
Kolmogorov-Smirnov	**D**	0.071732	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.035653	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.254649	**Pr > A-Sq**	>0.2500

Bris de glace

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	3.05	6	0.8020	-0.017	-0.124	-0.138	-0.071	-0.003	0.002
12	6.05	12	0.9133	-0.013	-0.068	0.080	0.019	0.000	-0.151
18	8.43	18	0.9716	0.035	0.020	0.099	0.011	0.056	-0.101

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.978817	**Pr < W**	0.2651
Kolmogorov-Smirnov	**D**	0.073024	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.072709	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.480128	**Pr > A-Sq**	0.2331

Vol partiel

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	7.26	6	0.2977	-0.017	-0.254	0.051	0.007	0.164	0.005
12	19.23	12	0.0832	-0.192	0.080	0.267	-0.117	-0.078	-0.088
18	25.26	18	0.1180	0.017	0.195	-0.105	-0.039	0.116	-0.012

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.970242	**Pr < W**	0.0856
Kolmogorov-Smirnov	**D**	0.080995	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.108003	**Pr > W-Sq**	0.0892
Anderson-Darling	**A-Sq**	0.703202	**Pr > A-Sq**	0.0670

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	5.87	6	0.4379	-0.174	0.014	0.014	-0.007	-0.087	0.192
12	17.08	12	0.1466	-0.044	0.113	0.124	-0.279	-0.042	-0.144
18	28.04	18	0.0614	0.022	0.092	0.176	-0.256	0.096	-0.030

Tests de normalité				
Test	Statistique		p Value	
Shapiro-Wilk	W	0.979005	Pr < W	0.2715
Kolmogorov-Smirnov	D	0.08085	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.050991	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.384373	Pr > A-Sq	>0.2500

Vol total

Tests sur les résidus des modèles UCM saturés sur les séries de coûts clos

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	6.66	6	0.3537	0.089	-0.075	-0.215	-0.151	-0.060	0.045
12	13.38	12	0.3420	0.128	-0.051	-0.205	0.068	0.122	-0.028

RC corporelle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	4.08	6	0.6653	-0.134	-0.046	-0.127	-0.063	0.000	0.118
12	11.04	12	0.5251	-0.034	-0.019	-0.128	0.178	0.081	0.161

RC matérielle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	3.83	6	0.6992	-0.135	-0.078	-0.032	-0.063	-0.025	-0.139
12	16.34	12	0.1762	-0.007	-0.091	0.068	0.122	-0.126	0.314
18	22.15	18	0.2256	-0.200	-0.032	-0.102	-0.018	0.068	-0.081

Dommages

Tests sur les résidus des modèles UCM saturés sur les séries de coûts déclarés

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	2.68	6	0.8480	-0.019	-0.098	-0.073	-0.127	-0.039	-0.049
12	7.31	12	0.8365	-0.066	-0.011	-0.002	0.121	-0.045	-0.181

RC matérielle

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	2.50	6	0.8683	-0.088	-0.088	0.000	-0.066	0.108	-0.035
12	5.39	12	0.9436	-0.053	0.043	-0.107	0.063	0.118	-0.007

Bris de glace

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	0.70	6	0.9945	0.046	-0.023	-0.054	-0.007	-0.038	-0.046
12	3.03	12	0.9953	-0.046	-0.021	-0.029	-0.029	-0.083	-0.125

Vol partiel

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	4.90	6	0.5563	-0.107	-0.226	-0.042	-0.033	0.013	0.008
12	9.04	12	0.6994	-0.035	-0.017	0.069	-0.095	-0.027	0.179

Vol total

Tests sur les résidus des modèles UCM non saturés (avec variables explicatives)

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	8.11	6	0.2301	-0.004	-0.020	0.084	-0.149	-0.070	-0.261
12	15.95	12	0.1936	-0.197	-0.043	-0.077	0.136	0.030	0.165

Tests de normalité

Test	Statistique		p Value	
Shapiro-Wilk	W	0.990646	Pr < W	0.8803
Kolmogorov-Smirnov	D	0.060506	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.033357	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.219892	Pr > A-Sq	>0.2500

Fréquences RC corporelle

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	3.46	6	0.7497	-0.076	0.174	-0.092	0.019	0.010	-0.033
12	8.06	12	0.7802	0.102	-0.044	0.018	0.021	-0.117	0.164

Tests de normalité

Test	Statistique		p Value	
Shapiro-Wilk	W	0.990859	Pr < W	0.8900
Kolmogorov-Smirnov	D	0.053135	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.029914	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.195522	Pr > A-Sq	>0.2500

Fréquences Bris de glace

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	16.38	6	0.0118	-0.264	0.298	-0.200	-0.000	-0.091	0.103
12	22.79	12	0.0295	-0.002	0.058	-0.153	0.198	0.043	0.087

Coûts clos RC matérielle

Autocorrelation Check for White Noise

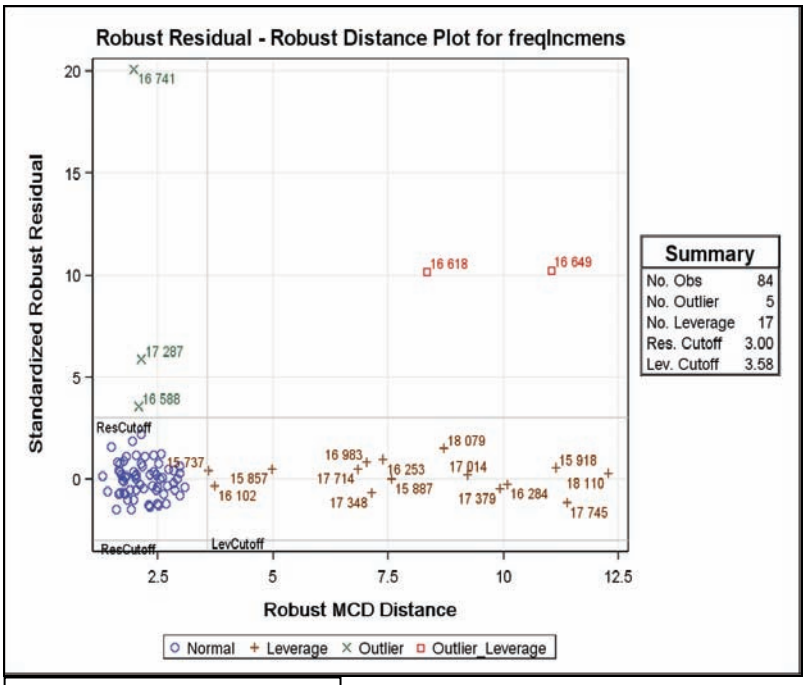
To Lag	Chi-Square	DF	Pr > Khi2	Auto-corrélations					
6	2.92	6	0.8189	-0.082	-0.072	-0.020	-0.085	0.115	0.072
12	5.88	12	0.9219	-0.030	0.043	-0.117	0.033	0.106	0.078

Coûts clos Bris de glace

ANNEXE 11 : Résultats de la régression linéaire (fréquences en Incendie)

Modèle de régression robuste avec la proc ROBUSTREG

Parameter Estimates							
Parameter	DF	Estimation	Erreur standard	95% Confidence Limits		Khi 2	Pr > Khi 2
FRT22	1	0.0059	0.0012	0.0036	0.0082	26.19	<.0001
FRT23	1	-0.0067	0.0027	-0.0120	-0.0014	6.17	0.0130
FRT24	1	-0.0093	0.0025	-0.0142	-0.0043	13.55	0.0002
FRT27	1	0.0018	0.0005	0.0008	0.0029	11.35	0.0008
Temperatures	1	0.0001	0.0000	0.0000	0.0002	9.69	0.0019
Scale	0	0.0012					



Profile for the Initial S Estimate	
Total Number of Observations	84
Number of Coefficients	5
Subset Size	5
Chi Function	Tukey
K0	2.9366
Breakdown Value	0.2500

Goodness-of-Fit	
Statistic	Value
R-Square	0.8052
AICR	69.8315
BICR	86.6945
Deviance	0.0001

MM Profile	
Chi Function	Tukey
K1	3.4400
Efficiency	0.8500

Diagnostics			
Obs	date	Standardized Robust Residual	Outlier
30	01JUN2005	3.5490	*
31	01JUL2005	10.1648	*
32	01AUG2005	10.2795	*
35	01NOV2005	20.0735	*
53	01MAY2007	5.8644	*

Diagnostics Summary		
Observation Type	Proportion	Cutoff
Outlier	0.0595	3.0000

Modèle de régression avec les variables d'intervention :

Parameter Estimates							
Parameter	**DF**	**Estimation**	**Erreur standard**	**95% Confidence Limits**		**Khi 2**	**Pr > Khi 2**
FRT22	1	0.0018	0.0004	0.0011	0.0026	22.46	<.0001
FRT27	1	0.0003	0.0001	0.0001	0.0005	7.86	0.0051
pulse1	1	0.0110	0.0008	0.0093	0.0126	168.92	<.0001
pulse2	1	0.0226	0.0011	0.0204	0.0247	425.47	<.0001
pulse3	1	0.0077	0.0011	0.0056	0.0099	49.85	<.0001
Scale	0	0.0011					

Autocorrelation Check for White Noise									
To Lag	**Chi-Square**	**DF**	**Pr > Khi2**	**Auto-corrélations**					
6	3.42	6	0.7545	0.020	-0.091	0.114	0.011	0.017	0.126
12	15.41	12	0.2199	0.016	-0.121	0.028	-0.202	0.017	0.255
18	20.99	18	0.2797	-0.038	-0.159	-0.127	0.056	0.011	-0.086

Tests de normalité				
Test	**Statistique**		**p Value**	
Shapiro-Wilk	**W**	0.964386	**Pr < W**	0.0201
Kolmogorov-Smirnov	**D**	0.067549	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.075124	**Pr > W-Sq**	0.2408
Anderson-Darling	**A-Sq**	0.524855	**Pr > A-Sq**	0.1844

ANNEXE 12 : Critères de comparaison et statistiques d'ajustement

Il existe plusieurs critères d'évaluation des performances d'un modèle. Ceux que nous avons retenus sont : le RMSE (Root Mean Squared Error), le MAPE (Mean Absolute Percentage Error), et le MPE (Maximum Percentage Error). En notant e_i les erreurs de prévision (avec $e_i = y_i - \hat{y}_i$), on a les formules suivantes :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \quad ; \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \; ; \; \text{MPE 1} = \max_{i=1}^{n} |e_i| \; ; \; \text{MPE 2} = \max_{i=1}^{n}\left|\frac{e_i}{y_i}\right|$$

Les statistiques d'ajustement dont nous avons tenu compte pour tester l'ajustement de nos modèles sont :

Le R carré : $R^2 = 1 - \dfrac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$, avec $\bar{y}$ la moyenne de la série

Le R carré ajusté : adjusted $R^2 = 1 - ([(n-1)/(n-k)])\,(1- R^2)$.

Le R carré ajusté d'Amemiya's : Amemiya's adjusted $R^2 = 1 - ([(n+k)/(n-k)])\,(1 - R^2)$.

Le critère que nous avons privilégié est le MAPE ; en effet, notre objectif est de minimiser l'erreur moyenne dans nos prévisions.