RÉALISER DES PRÉVISIONS PERTINENTES GRÂCE AU MODULE SAS® HIGH-PERFORMANCE FORECASTING (PARTIE I)

Caractéristiques :
Catégories : SAS® High-Performance Forecasting
OS : Windows, Unix, z/OS
Version : SAS® 9.3 et 9.2
Vérifié en juin 2012

Les procédures HPF (module SAS® High-Performance Forecasting) sont les procédures utilisées de manière automatique par la solution SAS® Forecast Studio, mais il est possible de les utiliser directement avec du code SAS pour le même résultat obtenu avec une personnalisation plus importante.
Cet article se compose de deux parties. La première concerne le fonctionnement général de ce type de procédures, la deuxième détaille les deux principales procédures. Dans un prochain article, nous expliquerons comment les autres procédures de ce module permettent une personnalisation des prévisions.
Le but de cet article est de vous montrer la puissance de ce module et des prévisions « industrielles » qu'il permet de réaliser.

Table des matières

Procédures HPF : le concept 1
PROC HPFDIAGNOSE 2
Description théorique 2
Un exemple d'utilisation de cette procédure 3
PROC HPFENGINE 5
Description de la procédure 5
Exemple d'utilisation de cette procédure 6
En cas de problème 9
Eléments à transmettre au Support Clients 9
Conclusion 9

Procédures HPF : le concept

Ces procédures ont pour but de générer des prévisions de grande qualité pour un large panel de types de modèles et de données. Elles peuvent être utilisées de manière automatique, via Forecast Studio, ou par un utilisateur averti, via du code, pour obtenir de meilleures prévisions de séries temporelles par rapport à SAS/ETS®. En effet, ces procédures offrent une large panoplie d'outils et de cas de figure supportés (événements, réconciliation,…).

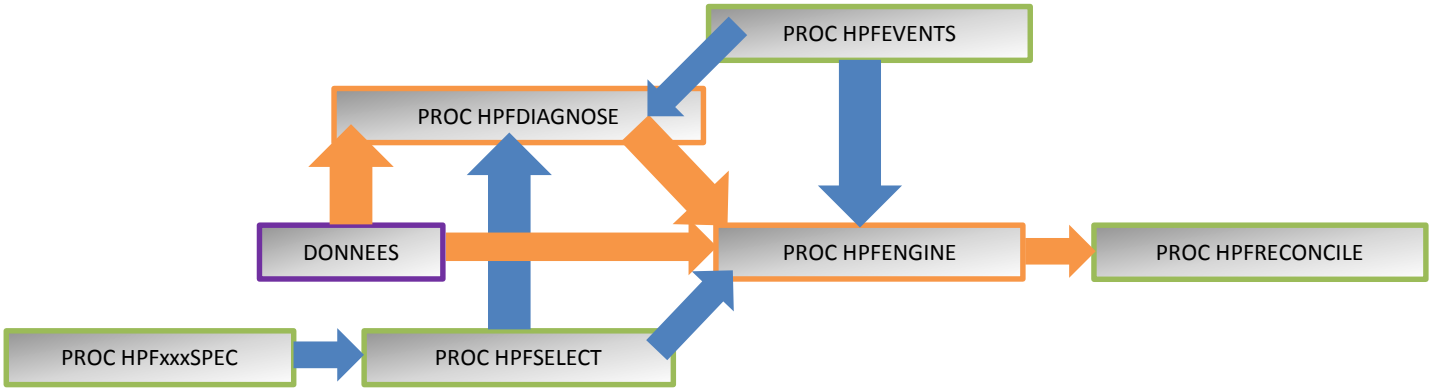
Chacune des procédures offre la possibilité de choisir une grande quantité d'options et de paramètres (qui sont déterminés ou choisis automatiquement lors de l'utilisation de la solution embarquée) pour personnaliser de manière très pointue et précise les différentes prévisions que nous voulons effectuer.

Le fonctionnement global de ces procédures est simple mais de nombreuses procédures sont disponibles si vous voulez une estimation personnalisée et précise. Des procédures permettent de créer des répertoires contenant des spécifications de modèle (PROC HPFxxxSPEC). Ensuite les meilleurs de chaque catégorie sont conservés (lissage, ARIMAX, demande intermittente et modèle à composantes inobservées). Enfin, on peut faire un diagnostic sur la série, l'estimer et la prévoir, en incorporant des options liées aux événements et aux réconciliations.

Ces procédures, ainsi que la solution, ont été conçues dans le but de pouvoir prévoir un grand nombre de séries de manière industrielle et ainsi traiter plus efficacement les grandes quantités de données qui sont désormais disponibles. De ce fait, comme nous le détaillerons dans les deux sections suivantes, la mise en production, c'est-à-dire la création des différentes spécifications de modèles, des toutes premières séries peut être assez longue et paraître fastidieuse. Cependant,

cette tâche permettra de prévoir rapidement et efficacement par la suite, puisque notre référentiel de modèles sera disponible et utilisable pour toutes les séries.

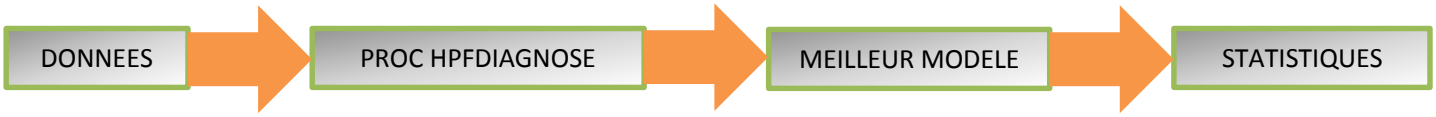
Voici une représentation globale des interactions entre les différentes procédures de ce module avec en orange le parcours des données et en bleu celui des spécifications :



Nous allons revenir plus en détail sur les procédures et leurs interactions dans les deux sections qui suivent.

PROC HPFDIAGNOSE

Après une description théorique du fonctionnement de la procédure, nous détaillerons un exemple de code et les résultats associés.



Description théorique

La procédure HPFDIAGNOSE fournit un ensemble complet d'outils pour l'identification de modèles de séries temporelles univariées de manière automatisée. Les données des séries temporelles peuvent avoir des outliers, des changements structurels et des effets de calendriers. Dans le passé, trouver un bon modèle pour les données de séries temporelles exigeait de l'expérience et de l'expertise en analyse de séries temporelles.

La procédure HPFDIAGNOSE diagnostique automatiquement les caractéristiques statistiques des séries temporelles et identifie les modèles appropriés. Elle considère pour chaque série temporelle les types de modèles suivants :
- autorégressif intégré avec moyenne mobile et inputs exogènes (ARIMAX),
- de lissage exponentiel, à demande intermittente, et à composantes non observées.
La transformation log et les tests de stationnarité sont effectués automatiquement. Les ordres autorégressifs (AR) et moyenne mobile (MA) sont déterminés, les outliers sont détectés et les meilleures variables d'inputs sont sélectionnées. Le diagnostic du modèle à composantes inobservées (UCM) trouve les meilleurs composants et sélectionne les meilleurs variables d'inputs.

La procédure HPFDIAGNOSE offre les fonctionnalités suivantes :

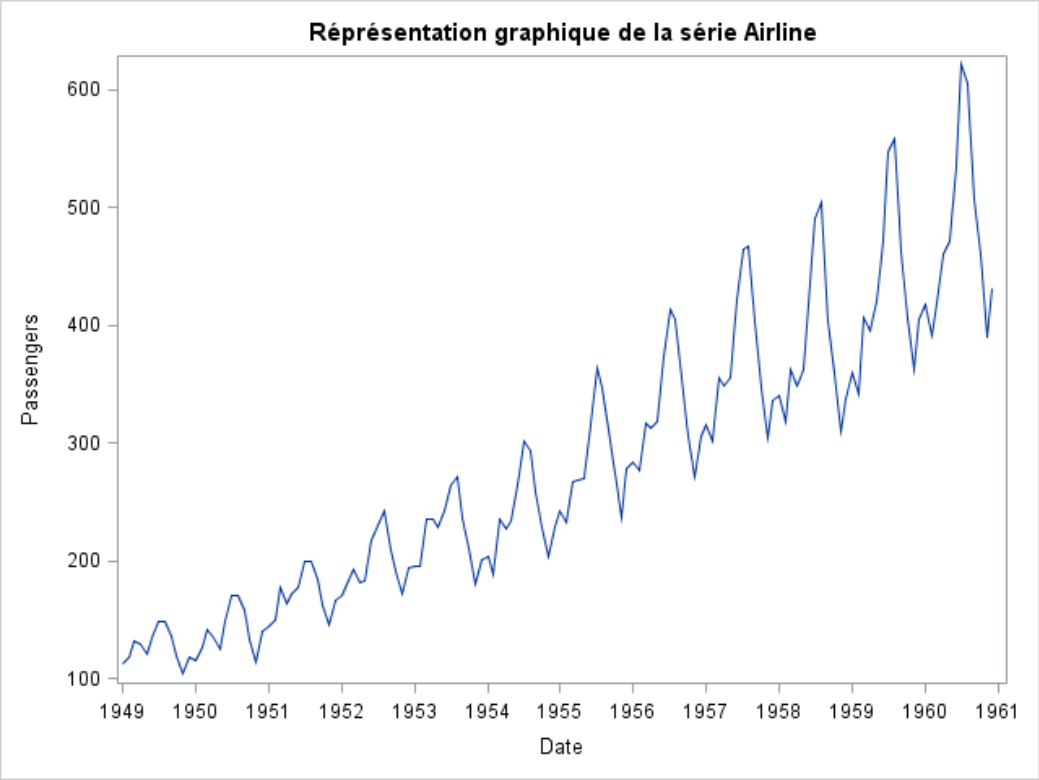
- Test d'intermittence (ou série interrompue)
- Test de transformation fonctionnelle
- Test de différenciation simple et saisonnière
- Ordre d'identification ARIMA simple provisoire
- Ordre d'identification ARIMA saisonnier provisoire
- Détection d'outliers
- Test de significativité des événements (variables indicatrices)
- Identification de la fonction de transfert
- Transformation fonctionnelle pour chaque régresseur

- Ordre de différenciation simple et ordre de différenciation saisonnière pour chaque régresseur
- Temps de retard pour chaque régresseur
- Ordres polynomiaux simples au numérateur et dénominateur pour chaque régresseur
- Modèle de demande intermittente (sélection automatique)
- Modèle de lissage exponentiel (sélection automatique)
- Modèle à composantes inobservées (sélection automatique).

On peut donc dire, pour résumer, que la procédure HPFDIAGNOSE sert à caractériser la série mais n'effectue ni prévision, ni estimation. La procédure nous fournit donc des statistiques sur la (ou les) série(s) étudiée(s) de la base de données. Il est possible de spécifier des sorties d'autres procédures (dont nous parlerons ultérieurement) en entrée de celle-ci. En effet, la procédure travaille sur un fichier de spécification par défaut (ce dont nous parlions auparavant au sujet de la première utilisation de cet outil) mais il est possible d'en créer un par vous-mêmes via ces autres procédures. On peut également utiliser les sorties de cette procédure, en particulier les modèles retenus, dans la procédure HPFENGINE.

Un exemple d'utilisation de cette procédure

Pour commencer, voici une représentation de la série sur laquelle nous allons travailler dans cet article



Voici un exemple de code que nous détaillerons:

```
/*PROC HPFDIAGNOSE*/
```

```
Proc hpfdiagnose criterion=mape data=tmp1.airline holdout=12
outest=estimateur outprocinfo=info outoutlier=outlier
repository=work.mesmodeles seasonality=12 specbase=maselection print=all;
arimax criterion=aic estmethod=ml method=minic identify=arima;
esm method=best;
forecast logpsngr passengers;
id date interval=month;
transform type=auto;
```

```
run;
```

Dans la première instruction, les options permettent respectivement de spécifier le critère utilisé par SAS (parmi une quarantaine possible), la table sur laquelle sera effectué le diagnostic, la taille de l'échantillon de validation, la table de sortie des estimateurs, la table de sortie des informations de cette procédure (celle où l'on trouvera les outliers de la série), le répertoire de référence pour la procédure concernant les modèles éligibles, la saisonnalité (si une saisonnalité est pressentie), un préfixe des modèles sélectionnés par la procédure et le type d'affichage des résultats.

L'instruction ARIMAX indique à SAS qu'il doit rechercher une spécification ARIMAX et les options présentées dans cet exemple permettent de sélectionner un critère (qui peut être différent du critère global), qui sera utilisé pour discriminer entre les différents modèles candidats de ce type. On choisit ensuite la méthode d'estimation (maximum de vraisemblance dans notre cas), la méthode de sélection des ordres AR et MA ainsi que l'ordre d'identification.

Concernant les lissages exponentiels, dans notre exemple, nous avons choisi de le laisser choisir le meilleur, au vu du critère MAPE, modèle possible, dans l'instruction ESM.

Ensuite nous lui indiquons les variables à expliquer, qui peuvent être multiples, comme dans notre exemple avec l'instruction FORECAST.

Avec ID, il faut par la suite lui indiquer la variable de temps ainsi que sa fréquence (ici mensuelle).

Enfin, avec l'instruction TRANSFORM, il est possible de réaliser des transformations si SAS identifie que cela améliore la prévision avec l'option TYPE= qui indique le type de transformation entre LOG, SQRT, LOGISTIC et BOXCOX pour ne citer que les principales. AUTO, le type de l'exemple permet un choix entre LOG et NONE effectué par la procédure.

Informations sur les variables	
Nom	Passengers
Libellé	Passengers
Premier(e)	JAN1949
Dernier(e)	DEC1960
Nombre d'obs. lues	144

Test de racine unité de Dickey-Fuller				
Type	Rho	Pr < Rho	Tau	Pr < Tau
Zero Mean	0.21	0.7309	1.38	0.9579
Single Mean	-2.32	0.7367	-1.05	0.7324
Trend	257.89	0.9999	-6.17	<.0001

Exemples 12 utilisés pour l'échantillon de validation

Test de racine unité de Dickey-Fuller saisonnier(Seasonality=12)				
Type	Rho	Pr < Rho	Tau	Pr < Tau
Zero Mean	-1.37	0.4451	-0.36	0.4044
Single Mean	-5.87	0.2523	-1.48	0.1305

Exemples 12 utilisés pour l'échantillon de validation

Test de racine unité jointe de Hasza-Fuller(Seasonality=12)					
		Valeurs critiques			
Type	Valeur F	90%	95%	99%	Approx. de Pr > F
Zero Mean	2.9637	2.5715	3.2631	4.8800	0.0640
Single Mean	3.1769	5.1415	6.3504	8.8400	0.1852
Trend	2.5461	7.2477	8.6677	10.7600	0.3295

Exemples 12 utilisés pour l'échantillon de validation

Critère d'information minimum						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	-6.24946	-6.32141	-6.30667	-6.32826	-6.29985	-6.27302
AR 1	-6.33466	-6.29602	-6.28026	-6.29538	-6.26474	-6.23915
AR 2	-6.32028	-6.28278	-6.25595	-6.25609	-6.22534	-6.2063
AR 3	-6.3503	-6.3141	-6.27621	-6.24523	-6.24194	-6.22247
AR 4	-6.33057	-6.29054	-6.25187	-6.25848	-6.21998	-6.1958
AR 5	-6.30796	-6.26784	-6.22782	-6.23527	-6.19898	-6.165

Exemples 12 utilisés pour l'échantillon de validation

Test de transformation fonctionnelle	
Variable	Transformation fonctionnelle
Passengers	LOG

Synthèse du test de racine unité de Dickey-Fuller				
Variable	Saisonnalité	Moyenne zéro	Moyenne	Tendance
Passengers	1	YES	YES	NO

Synthèse du test de racine unité de Dickey-Fuller saisonnier				
Variable	Saisonnalité	Moyenne zéro	Moyenne	Tendance
Passengers	12	YES	YES	

Synthèse du test de racine unité jointe				
Variable	Saisonnalité	Moyenne zéro	Moyenne	Tendance
Passengers	1, 12	YES	YES	

Spécification du modèle ARIMA													
Variable	Transformation fonctionnelle	Constante	p	d	q	P	D	Q	Saisonnalité	Critère du modèle	Statistique	Echantillon de validation	Statut
Passengers	LOG	NO	1	1	0	0	1	1	12	MAPE	3.4374	12	OK

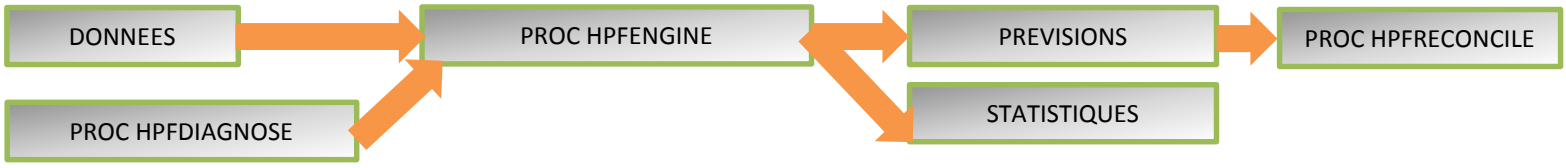
Spécification du modèle de lissage exponentiel						
Variable	Transformation fonctionnelle	Modèle, sélectionné	Composante	Critère du modèle	Statistique	Echantillon de validation
Passengers	LOG	ADDWINTERS	LEVEL	MAPE	3.5571	12
			TREND			
			SEASONAL			

Spécification du meilleur modèle													
Variable	Transformation fonctionnelle	Constante	p	d	q	P	D	Q	Saisonnalité	Critère du modèle	Statistique	Echantillon de validation	Statut
Passengers	LOG	NO	1	1	0	0	1	1	12	MAPE	3.4374	12	OK

Dans les résultats, on retrouve diverses informations, que l'on a évoquées pour la plupart dans les options, qui correspondent bien à ce que l'on a demandé. Ainsi, une transformation fonctionnelle a été effectuée (logarithmique en l'occurrence). Le choix s'effectue entre le meilleur modèle ARIMA (qui a été détecté par la méthode MINIC avec comme critère AIC, même si le seul indiqué est la MAPE, pour que l'on puisse faire des comparaisons entre les types de modèles) et le meilleur lissage (dont les composantes utilisées sont listées). La procédure HPFDIAGNOSE choisit, au vu du critère spécifié, comme meilleur modèle prédictif de cette série, le modèle ARIMA(1,1,0)x(0,1,1).

PROC HPFENGINE

Cette nouvelle partie traite de la procédure HPFENGINE dédiée aux prévisions.



Description de la procédure

La procédure HPFENGINE fournit une manière automatique de générer des prévisions pour plusieurs données de séries temporelles ou transactionnelles en une seule étape. La procédure peut automatiquement choisir le meilleur modèle de prévision à partir d'une liste de modèles définie par l'utilisateur ou à partir d'une liste de modèles par défaut. Vous pouvez générer des spécifications ou

les choisir à partir d'un ensemble par défaut. Les familles de modèle supportées incluent encore une fois le lissage exponentiel, la demande intermittente, la composante inobservée et ARIMA.

Tous les paramètres associés au modèle de prévision sont optimisés sur la base des données. La procédure HPFENGINE sélectionne le modèle approprié pour chaque série de données sur la base d'un des nombreux critères de sélection de modèle (MAPE, RMSE,…).

La procédure fonctionne dans une variété de modes. Lors de son usage le plus complet, tous les modèles candidats appropriés à partir d'une liste des modèles concernés sont aptes pour une série particulière, et le modèle qui produit le meilleur ajustement (basé sur un critère défini par l'utilisateur) est déterminé. Les prévisions sont ensuite produites à partir du modèle. Il est également possible d'ignorer le processus de sélection, d'ajuster un modèle particulier et de produire des prévisions. Enfin, étant donné un ensemble d'estimations de paramètres et de spécifications de modèles à partir des résultats de la procédure HPFDIAGNOSE, la procédure peut contourner entièrement l'étape d'ajustement et calculer directement les prévisions.

La procédure HPFENGINE écrit la série temporelle extrapolée par les prévisions, les statistiques sommaires des séries, les prévisions et limites de confiance, les paramètres estimés et les statistiques d'ajustement de l'ensemble dans des tables de sortie.

Elle peut prévoir à la fois des données de séries temporelles (dont les observations sont « equally spaced » à un intervalle de temps spécifique) et des données transactionnelles (dont les données ne sont pas espacées par un intervalle de temps donné). Pour les données transactionnelles, les données sont accumulées à un intervalle de temps spécifié pour former une série temporelle (par exemple, on additionne les données de ventes d'un même mois pour obtenir le chiffre d'affaire du mois en question).

En résumé, cette procédure est celle parmi cette gamme qui estime et qui prévoit les séries. Elle offre un large panel de possibilités, d'options de prévisions grâce à différents outils. En effet, comme pour la procédure HPFDIAGNOSE, il est possible d'utiliser les sorties d'autres procédures HPF, notamment HPFDIAGNOSE. Enfin, les résultats obtenus peuvent être réintroduits dans la PROC HPFRECONCILE dont nous parlerons dans la section suivante.

Exemple d'utilisation de cette procédure

Nous allons commencer par expliquer le code utilisé avant de passer aux résultats :

```
/*PROC HPFENGINE*/

ods graphics on;
Proc hpfengine data=tmp1.airline globalselection=best inest=estimateur
                    out=eng1 outcomponent=eng2 outest=eng3 outfor=eng4
outindep=eng5
                    outmodelinfo=eng6 outprocinfo=eng7 outstat=eng8
outstatselect=eng9
                    plot=all print=all printdetails
repository=work.mesmodeles
                    seasonality=12 scorerepository=work.score
task=select(override holdout=12 criterion=mape);
forecast logpsngr passengers;
id date interval=month;
score;
run;
ods graphics off;
```

Nous ajoutons les instructions ODS GRAPHICS ON/ODS GRAPHICS OFF pour profiter de toutes les sorties graphiques de cette procédure.
Instruction PROC HPFENGINE, voici les mots clé les plus intéressants :

- GLOBALSELECTION : spécifie le nom du catalogue qui sert en entrée de la liste de la sélection de modèle, BEST est l'option par défaut. Option facultative si INEST= est renseignée.

- INEST : table d'entrée contenant les cartes d'informations issues de la procédure HPFDIAGNOSE concernant les variables à expliquer et les estimateurs.
- OUT : table de sortie avec les prévisions effectuées ajoutées à la fin de la table initiale.
- OUTCOMPONENT : table contenant les composantes prévues pour les lissages.
- OUTEST : table de sortie contenant les informations sur les estimateurs. Cette table est beaucoup plus complète que celle issue de la procédure HPFDIAGNOSE.
- OUTFOR : table où l'on retrouve toutes les prévisions du modèle, pour les données connues et futures, avec les intervalles de confiance et les erreurs de prévision.
- OUTINDEP : table qui ne se génère qui si une des variables inputs est stochastique. Elle contient les prévisions pour ce type de variable, ce qui n'est pas le cas dans notre exemple.
- OUTMODELINFO : table contenant l'information sur les modèles (saisonnier ?, avec un trend ?;…).
- OUTPRCINFO : table où est répertoriée de l'information sur le déroulement de la procédure (nombre d'erreurs, de notes, de warnings,…).
- OUTSTAT : la valeur des différentes statistiques calculées par cette procédure est reprise dans cette table pour chacune des séries prévues. Il s'agit d'une des tables de sorties les plus importantes.
- OUTSTATSELECT : dans cette table, on retrouve les statistiques décrites ci-dessus, pour chacun des modèles candidats en entrée de la procédure et cela pour toutes les séries prévues.
- SCOREREPOSITORY : spécifie l'emplacement du catalogue où est repris le modèle de score correspondant au modèle. Nécessite la présence de l'instruction SCORE.
- TASK : cette option contrôle la méthode de sélection du modèle champion. Dans notre exemple, SELECT ignore le contenu de la table INEST et le choisit avec un échantillon de validation de 12, suivant la MAPE. La documentation en ligne reprend en détail la liste des options possibles.

Les autres instructions possèdent une structure identique à celles de la procédure HPFDIAGNOSE avec, dans l'instruction FORECAST, la liste des variables à prévoir et ID qui reprend l'identifiant temporelle et ses caractéristiques. L'instruction SCORE permet le calcul d'un modèle de SCORE (nécessite la présence de l'option SCOREREPOSITORY dans l'instruction principale).

Les résultats partiels obtenus avec ce programme :

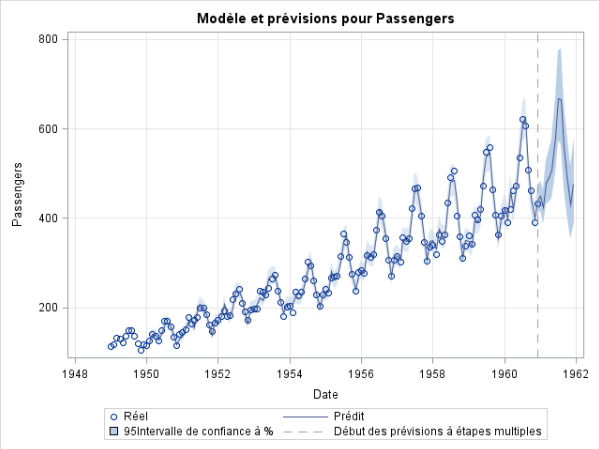
Informations sur les variables	
Nom	Passengers
Libellé	Passengers
Premier(e)	JAN1949
Dernier(e)	DEC1960
Nombre d'observations lues	144

Statistiques descriptives	
Variable	Passengers
Nombre d'observations	144
Nombre d'observations manquantes	0
Minimum	104
Maximum	622
Moyenne	280.2986
Ecart-type	119.9663

Critère de sélection du modèle = MAPE			
Modèle	Statistique	Sélectionné	Libellé
MASELECTION2	3.4373944	Oui	ARIMA: Log(Passengers) ~ P = 1 D = (1,12) Q = (12) NOINT
MASELECTION3	3.5571486	Non	Log Winters Method (Additive)

Valeurs estimées des paramètres pour le modèle MASELECTION2					
Composante	Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx Pr > |t|
Passengers	MA1_12	0.5618	[illegible]	6.59	<.0001
Passengers	AR1_1	-0.33953	0.08214	-4.13	<.0001

Graphique du modèle



Statistiques d'ajustement pour Passengers	
Statistique	Valeur
Erreur de degré de liberté	129
Nombre d'observations	144
Nombre d'obs. utilisées	131
Nombre de valeurs réelles manquantes	0
Nombre de valeurs prédites manquantes	13
Nombre de paramètres du modèle	2
Somme totale des carrés	13163908
Somme totale corrigée des carrés	1714610.41
Somme de l'erreur carrée	15478.1312
Carré moyen de l'erreur	118.153673
Racine carrée du carré moyen de l'erreur	10.8698516
Carré moyen de l'erreur sans biais	119.985513
Racine carrée de la moyenne du carré des erreurs sans biais	10.9537899
Erreur relative (en) de la moyenne absolue	2.97393995
Erreur absolue de la moyenne	8.33478443
R carré	0.9909728
R carré ajusté	0.99090282
R carré ajusté d'Amemiya	0.99069289
R carré de cheminement aléatoire	0.90338481
Critère d'information d'Akaike	629.130178
Critère d'information bayésien de Schwarz	634.880572
Critère de prévision d'Amemiya	121.817353
Erreur max.	40.3335794
Erreur min.	-40.890711

Synthèse des prévisions													
Variable	Valeur	JAN1961	FEB1961	MAR1961	APR1961	MAY1961	JUN1961	JUL1961	AUG1961	SEP1961	OCT1961	NOV1961	DEC1961
LogPsngr	Réel(le)	.	.	.	.	.	.	.	.	.	.	.	.
LogPsngr	Prédit(e)	6.10553	6.04881	6.16730	6.19425	6.22745	6.36381	6.50211	6.49782	6.31981	6.20393	6.05857	6.16318
LogPsngr	Erreur type	0.03726	0.04466	0.05320	0.05986	0.06606	0.07167	0.07688	0.08176	0.08636	0.09073	0.09490	0.09889
LogPsngr	Borne inférieure	6.03249	5.96128	6.06303	6.07692	6.09797	6.22335	6.35143	6.33758	6.15054	6.02610	5.87257	5.96935
LogPsngr	Borne supérieure	6.17857	6.13634	6.27156	6.31158	6.35694	6.50428	6.65280	6.65806	6.48907	6.38176	6.24457	6.35701
LogPsngr	Erreur	.	.	.	.	.	.	.	.	.	.	.	.
Passengers	Réel(le)	.	.	.	.	.	.	.	.	.	.	.	.
Passengers	Prédit(e)	448.6413	424.0299	477.5703	490.8007	507.5709	581.9487	668.5200	665.9138	557.5400	496.7315	429.6949	477.2659
Passengers	Erreur type	16.72427	18.94612	25.42404	29.40706	33.56883	41.75948	51.47263	54.53478	48.24002	45.16214	40.87031	47.31440
Passengers	Borne inférieure	416.7523	388.1055	429.6750	435.6840	444.9540	504.3910	573.3101	565.4238	468.9702	414.0981	355.1607	391.2532
Passengers	Borne supérieure	482.3002	462.3566	529.3044	550.9121	576.4781	667.9924	774.9475	779.0385	657.9118	590.9693	515.2098	576.5215
Passengers	Erreur	.	.	.	.	.	.	.	.	.	.	.	.

Concernant les résultats, on trouve, sur la capture d'écran qui commence par les informations sur les variables, la liste des modèles utilisés. Le graphique nous illustre la vraisemblance du modèle retenu. Le tableau sur les statistiques d'ajustement nous informe sur la qualité de l'ajustement (à comparer à d'autres modèles ou techniques d'estimations). La synthèse a été reprise pour illustrer la puissance de cette procédure capable de prévoir plusieurs séries en simultané.

En cas de problème

Eléments à transmettre au Support Clients

Si vous rencontrez des problèmes lors de l'utilisation de ce module, vous pouvez nous écrire à support@sas.com, en attachant à votre message l'erreur reçue dans votre journal.

Conclusion

Le module SAS® High-Performance Forecasting est LE module SAS pour faire des prévisions de qualité, dans un mode industriel. Il supporte la plupart des méthodes de prévisions (ARIMA, ESM, IDM, UCM). Nous avons dans cette article présenté les deux principales procédures de ce module. Dans un article, à paraître le trimestre prochain, seront détaillées les autres procédures de ce module disponibles. A suivre…

Jérémy NOEL
Consultant Support Clients SAS France