A la Une du Support

TIMES SERIES DATA MINING : NOUVELLES PERSPECTIVES

Caractéristiques :

Catégories : SAS Enterprise Miner
OS : Windows
Versions : 7.1 et 12.1
Vérifié en juin 2013

La version 12.1 de SAS® Enterprise Miner™ voit l'arrivée en production de la famille des nœuds de Time Series Data Mining (TSDM). Il s'agit d'une réelle innovation en soi, permettant de réconcilier l'étude des séries temporelles et le Data Mining avec leurs spécificités respectives et qui étaient jusqu'à présent antagonistes. Découvrons ensemble trois de ces nœuds présents sur ce nouvel onglet TSDM (Séries chronologiques pour la version française) : « Time Series Data Preparation », « TS Similarity » et « TS Exponential Smoothing ». De prime abord, cet article vous permettra de découvrir les méthodologies associées, comme l'analyse de similarité, avant de détailler chacun des nœuds qui sont pour le moment disponibles, à travers quelques exemples simples.

Sommaire

1. Un contexte particulier 1
2. L'analyse de similarité 2
   2.1. Généralités 2
   2.2. Exemple de séquences temporelles 3
3. Le Time Series Data Mining avec SAS Enterprise Miner 4
   3.1. « Time Series Data Preparation » 4
   3.2. « Time Series Similarity » : construction d'une matrice de similarité 6
   3.3. « Time Series Exponential Smoothing » 8
4. En cas de problème 9
   4.1. Les problèmes les plus fréquents 9
   4.2. Contacter le Support Clients 9
5. Liens utiles 9
6. Conclusion 10

1. UN CONTEXTE PARTICULIER

Les applications transactionnelles collectent une grande quantité de données horodatées relatives à la consommation des clients, des fournisseurs au cours du temps. De ce fait, une entreprise peut avoir à gérer une quantité considérable de clients, de fournisseurs avec un ensemble de transactions associées à chacun d'entre eux.

Ainsi on passe de données présentées en lignes ou sous forme transactionnelle à des données directement exploitables. La chose essentielle à retenir de cette section, et qui est valable en dehors de ce sujet précis, est la différence entre la donnée transactionnelle (« là où naît l'information ») et la donnée consolidée (retravaillée, épurée,… ex : DataWarehouse ou non pour l'apprentissage supervisé). Le TSDM permet d'étudier des données chronologiques de n'importe quel type en les transformant aisément en données exploitables et « parlantes ».

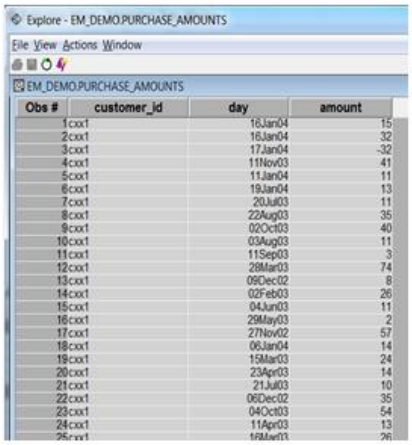
• apprentissage supervisé • séries temporelles

ID	Age	Revenue	Products	Churn
12	23	1234	3	0
13	45	432	1	1

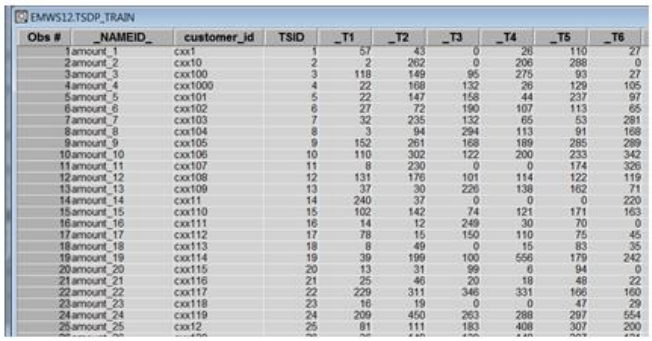
Time	ID_12	ID_13
Mar08	23	67
Apr08	43	58
May08	37	51

ID	Age	Revenue	Products	M_{0-3}	M_{0-2}	M_{0-1}	Churn
12	23	1234	3	23	43	37	0
13	45	432	1	67	58	51	1

Sur un exemple plus complexe, on obtient un résultat du type :



- à partir de données transactionnelles longues et étroites
- une observation par transaction
- plusieurs observations par ID client



- dans des vues clients courtes et larges
- une observation par ID client
- une variable par période temporelle et par combinaison cross-ID

Nous exposerons certaines techniques d'analyse de similarité en utilisant les mesures de similarité, combinées avec l'analyse et/ou la décomposition automatique des séries temporelles. Conséquemment à l'analyse transactionnelle, nous utiliserons des techniques d'apprentissage NON supervisé et autres techniques de data mining traditionnelles.

2. L'ANALYSE DE SIMILARITE

2.1. Généralités

Prenons deux séquences numériques ordonnées (que nous nommerons respectivement input et cible), telles que deux séries temporelles. Une mesure de similarité est une métrique qui mesure la distance entre la séquence input et la séquence cible tout en tenant compte de l'ordonnancement. Les mesures de similarité peuvent être calculées entre la séquence input et la séquence cible si bien que des mesures de similarité qui « glissent » (déformation) sur la séquence cible au regard de la séquence d'input peuvent être obtenues. Les « glissements » peuvent se produire par index de l'observation (mesures de similarité par séquences coulissantes) ou par index saisonnier (mesures de similarité par séquences saisonnières coulissantes).
En calculant la mesure de similarité entre les deux séries temporelles, vous avez besoin de tâches pour préparer les données des séries temporelles. Ces tâches vont transformer les séries temporelles, normaliser les séquences, mettre à l'échelle les séquences et calculer des métriques ou des mesures.

Pour préparer les données des séries temporelles, vous pouvez utiliser le lissage, la décomposition/ajustement et d'autres méthodes. Pour transformer les séries temporelles cible et d'input, vous pouvez utiliser des fonctions comme la différenciation (simple et saisonnière), les

transformations et les autres fonctions de transformations. Il est aussi possible de normaliser, de mettre à l'échelle ou de calculer les métriques des séquences cible et d'input.

Les mesures de similarité peuvent être utilisées pour comparer une séquence cible unique à plusieurs autres séquences d'input. Cette situation se pose en recherche et recouvrement de séries temporelles ou en classement. Par exemple, étant donnée une séquence cible unique, vous pouvez trouver des séquences d'input qui sont « similaires » ou « proches » de la cible. Les recherches et recouvrements et les analogies sont importants pour les prévisions des nouveaux produits et les prévisions de séries temporelles analogues. Ces techniques sont utiles lorsqu'un grand nombre de séries temporelles historiques sont disponibles.

Les mesures de similarité peuvent être utilisées pour comparer une unique séquence d'input à plusieurs autres séquences cibles représentatives. Cette situation se pose en classification de séries temporelles. Par exemple, étant donné une unique séquence d'input, vous pouvez classifier la séquence d'input en trouvant la « plus similaire » ou la « plus proche » séquence cible. Cette analyse peut être répétée pour classifier un grand nombre de séquences d'input.

Les mesures de similarité peuvent être calculées entre plusieurs séquences pour former une matrice de similarité. Par exemple, étant donné K séquences temporelles, une (KxK) matrice de symétrique peut être construite dont le *ij*ième élément contient la mesure de similarité entre la *i*ième et la *j*ième séquence. Les techniques de clustering peuvent ensuite être appliquées à la matrice de similarité. Cette situation se pose en clustering de séries temporelles. Il permet ainsi de regrouper des produits en « groupe » ou « famille » de produits qui suivent un cycle de vie similaire ou de former des groupes d'actions qui suivent un comportement identique face à la même situation.

2.2. Exemple de séquences temporelles

Soit X = {3,5,3,3,3,6,3,8} le vecteur de la séquence d'input avec N_x= 8. Soit Y = {2,4,6,7,8,9,3,10,11} le vecteur de la séquence cible avec N_y= 10. Les séquences cible et d'input sont représentées sur les figures 1 et 2.

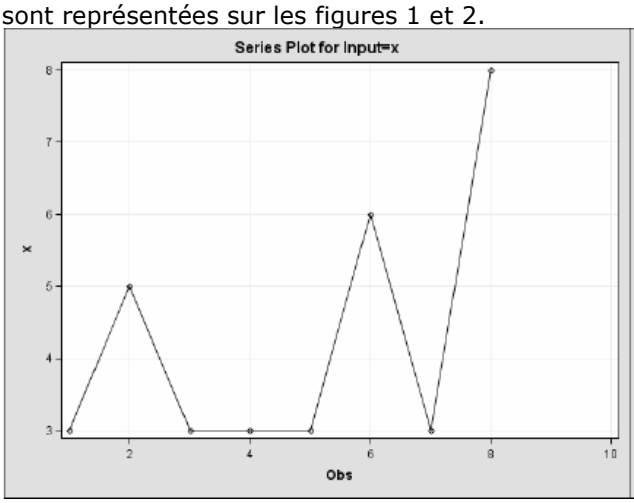
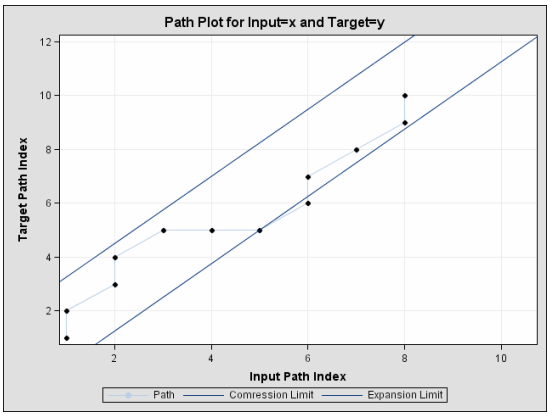


Figure 1 : Graphique de la séquence d'input

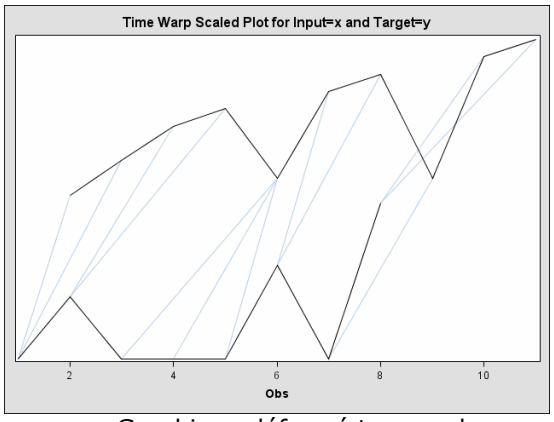
Figure 2 : Graphique de la séquence cible

Notez que les graphiques des deux séquences temporelles sont assez similaires mais restent différents en longueur et en timing. Chaque séquence a deux phases de mouvements ascendants, qui sont suivies par des mouvements descendants et une phase finale de mouvement ascendant.

A partir de là, nous allons étirer, réduire des portions de ces deux séquences pour les faire correspondre au mieux en utilisant le moins de déformations.
Ensuite, il faut créer un chemin (nom de la façon qu'on utilise pour rapprocher deux séries distinctes) représenté ci-dessous sur le graphique de gauche pour obtenir le graphique de droite qui nous explicite comment rapprocher les deux séries temporelles en déformant les phases ascendantes et descendantes (ces techniques se nomment compression et expansion de la série temporelle).



Graphique de chemin



Graphique déformé temporel

Les deux séquences chemin ont deux phases ascendantes qui sont suivies par des mouvements descendants et une phase finale de mouvement ascendant. Notez que les deux graphiques des séquences temporelles sont assez semblables, mais à la différence des graphiques dans les figures 1 et 2, ils ont maintenant la même longueur et le même timing (déformé).

Cela permet de définir des métriques de distance et de similarité, qui seront utilisées dans les nœuds du produit pour permettre de définir les séries qui sont les plus proches et celles qui sont les plus éloignées en terme de comportement.

La similarité peut donc être résumée comme suit :
L'utilisation possible des mesures directes entre Séries Temporelles (séquences) :
Mesures de Distance observant strictement à la fois les notions d'ordonnancement (*ordering*) et de *timing*

- PROC DISTANCE (SAS/STAT)

Mesures de Similarité observant strictement l'ordonnancement mais approximativement le timing

- PROC SIMILARITY (SAS/ETS)
  - sliding : décalage dans le temps de l'occurrence des motifs
  - warping : déformation du signal du motif

Cette partie théorique est très succincte mais elle est essentielle pour mieux comprendre ce qui est réalisé par SAS Enterprise Miner. A la fin du document vous trouverez un lien vers un papier international, écrit par un développeur du logiciel SAS, expert dans ce domaine si particulier, qui donne plus de détails sur l'analyse de similarité.

3. Le Time Series Data Mining avec SAS Enterprise Miner

Le TSDM, méthode récente, permet d'analyser les séries temporelles en tenant compte des particularités de ce type de données en gardant les avantages des techniques de Data Mining. Cette méthodologie a été introduite de manière expérimentale avec la version 7.1 de SAS Enterprise Miner. Avec la version 12.1, les trois nœuds principaux ont été complètement intégrés. Il est possible que d'autres nœuds, testés de manière expérimentale avec la version 7.1, soient également proposés dans le futur.

3.1. « Time Series Data Preparation »

La préparation de données en analyse prédictive se base sur :

- La visualisation de données TS
- La création de variables additionnelles (input) pour l'analyse prédictive :
  - Il y a potentiellement un grand nombre de variables possibles
  - Il est nécessaire (ou recommandé ?) de compresser tout en minimisant la perte d'information
    - longitudinale
    - séries temporelles
- L'agrégation de l'information se fonde sur :
  - la décomposition des séries temporelles
  - la réduction de dimension

- l'analyse de similarité
- La combinaison des séries temporelles basée sur le clustering

En séries temporelles, nous avons en général des données transactionnelles (pour une date donnée, des informations en fonction d'autres variables mais toujours avec le temps comme clé) qu'il va falloir changer.
Par exemple, trois années de données de ventes de cosmétiques en fonctions de 3 variables : l'état (5 modalités), le groupe de produits (5 différents) et le groupe de consommateurs (5 aussi). Vous avez ci-dessous, la table et le paramétrage dans SAS Enterprise Miner.

EMWS7.Ids_DATA

	Time ID	Target	CrossID: product	CrossID: group	CrossID: state
1	Jan 1, 1996	82971.0	54105	A	NC
2	Jan 1, 1996	82322.0	54105	A	GA
3	Jan 1, 1996	95834.0	54105	A	WI
4	Jan 1, 1996	98710.0	54105	A	MD
5	Jan 1, 1996	123089.0	54105	A	FL
6	Jan 1, 1996	86419.0	54105	B	NC
7	Jan 1, 1996	81668.0	54105	B	GA
8	Jan 1, 1996	75308.0	54105	B	WI
9	Jan 1, 1996	84005.0	54105	B	MD
10	Jan 1, 1996	91913.0	54105	B	FL
11	Jan 1, 1996	95748.0	54105	C	NC
12	Jan 1, 1996	99522.0	54105	C	GA
13	Jan 1, 1996	107240.0	54105	C	WI
14	Jan 1, 1996	102089.0	54105	C	MD
15	Jan 1, 1996	124963.0	54105	C	FL
16	Feb 1, 1996	83338.0	54105	A	NC
17	Feb 1, 1996	71453.0	54105	A	GA
18	Feb 1, 1996	75220.0	54105	A	WI

Variables - Ids

(none) not Equal to

Columns: Label

Name	Role	Level
MNTH_YR	Time ID	Interval
SALES	Target	Interval
SKU	Cross ID	Nominal
group	Cross ID	Nominal
state	Cross ID	Nominal

Il faut maintenant changer cette table en une table exploitable pour les autres nœuds du TSDM.

Property	Value
General	
Node ID	TSDP
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Time Interval	
Specify an Interval	Month
Seasonal Cycle Selection	Default
Length of Cycle	2
Start Time	DEFAULT
End Time	DEFAULT
Time of Day	Default
Accumulation	Total
Transform	
Transformation	None
Box-Cox Parameter	0.0
Difference	
Apply Differencing	No
Difference Order	1
Seasonal Differencing	No
Missing Value	
Set Value	Constant
Constant Value for Missing Observation	0.0
Transpose	
Transpose	Yes
By Variable	By TimeID
ODS Graphs	None
Status	
Create Time	2/3/11 2:30 PM
Run Id	8246d93c-3783-4059-a486-bed8ba775
Last Error	

Ci-dessous, les paramètres disponibles et les fonctionnalités de ce nœud (illustré ci-contre avec la liste des options de celui-ci):

- crée un ID de séries temporelles
- crée des métadonnées de séries temporelles
- transpose les données de séries temporelles
- détecte et spécifie l'intervalle de temps, la saisonnalité, les dates de début et de fin
- crée des séries temporelles à partir de données transactionnelles par agrégation
- valeurs manquantes, différenciation, transformation, etc.

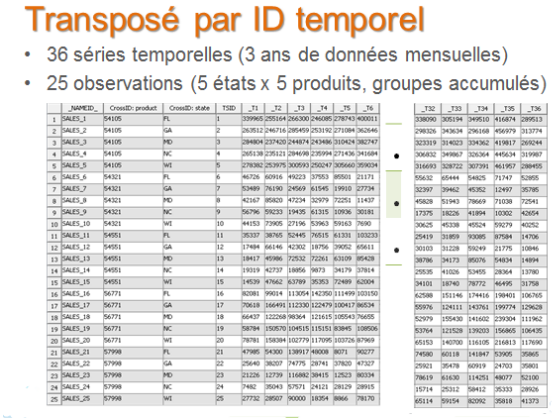
De plus, il est donc possible de transposer les données selon deux axes :

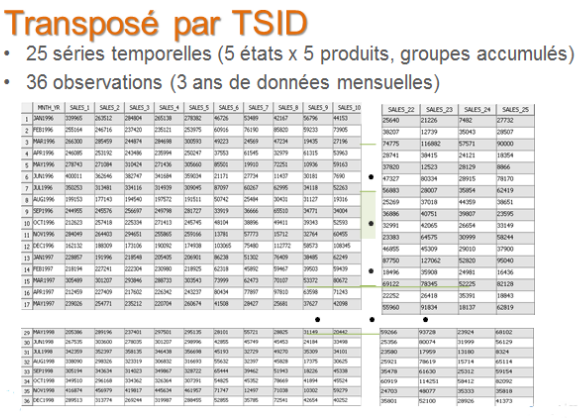
- par ID temporel

Les Cross-ID deviennent des observations et les intervalles de temps deviennent des variables.

- par TS ID
  Les intervalles de temps deviennent des observations et les Cross-ID deviennent des variables.

Dans notre exemple, cela donne :



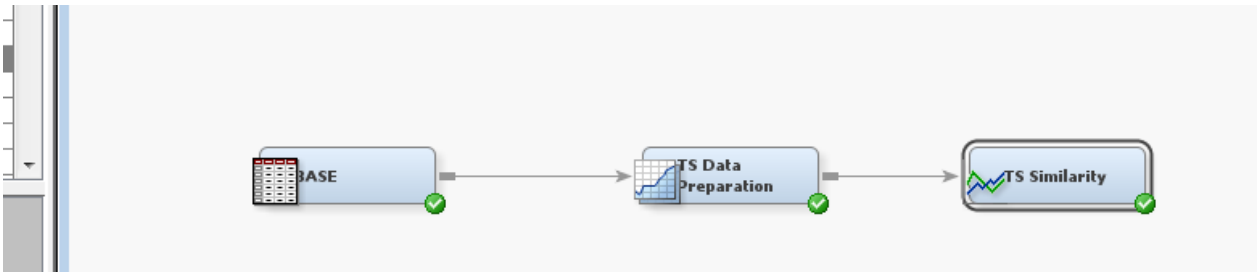


Ce premier nœud permet donc aux utilisateurs de manipuler des données de séries temporelles et transactionnelles pour faciliter le TSDM. Ce nouveau nœud fournit plusieurs types d'outils de manipulation des données des séries chronologiques, incluant les définitions des intervalles de temps, les transpositions et transformations de données, la différenciation des données et les affectations des valeurs manquantes. Nous pouvons donc choisir l'axe d'étude, soit nous considérons les séries temporelles comme des observations pour faire du clustering sur celles-ci (Transpose by TimeID dans le nœud), soit comme des variables (Transpose by TSID) pour faire une analyse de similarité. De même, les possibilités de transformations et d'imputations sont assez nombreuses et variées.

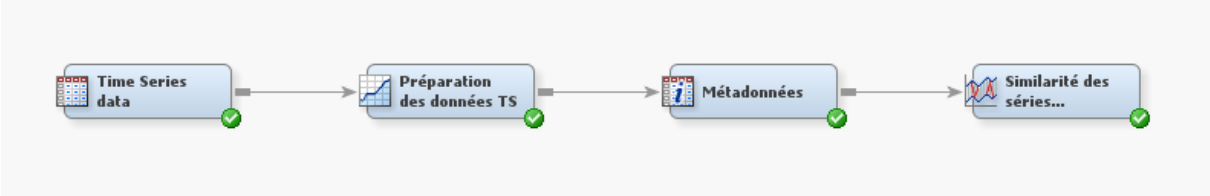
3.2. « Time Series Similarity » : construction d'une matrice de similarité

La première chose que nous allons montrer est une « similarity map » qui représente les similarités qui existent entre les différentes variables. Le bleu montre une mesure de similarité forte et le rouge indique à l'inverse une faible similarité. Plus la couleur est vive, plus la mesure est intense (fortement ou faiblement), le blanc étant la neutralité. Nous n'allons pas détailler la construction de cette carte, qui reprend l'analyse et les mesures de similarité que nous avons montrées dans la section précédente, mais seulement expliquer, à travers un exemple, les résultats obtenus. L'exemple sur lequel nous allons travailler est le CAC40. Les données du CAC40 ont été extraites du site de Yahoo ! Finance grâce au programme disponible dans le livre de monsieur Sébastien Ringuédé, SAS 2e édition Introduction pratique : du data management au reporting. Ainsi, au chapitre 10 se trouve un exemple de code permettant l'extraction des données du site et le retraitement dans une table SAS. En effet, nous disposons d'une table contenant les données des 40 titres et l'indice à la fermeture tous les jours entre le 1er septembre 2009 et le 31 janvier 2011. Des valeurs manquantes sont présentes en début de période pour un titre, ce qui va nous permettre d'étudier le comportement de la solution dans cette situation. Le nœud « Time Serie Similarity » est celui permettant la création de la carte de similarité.

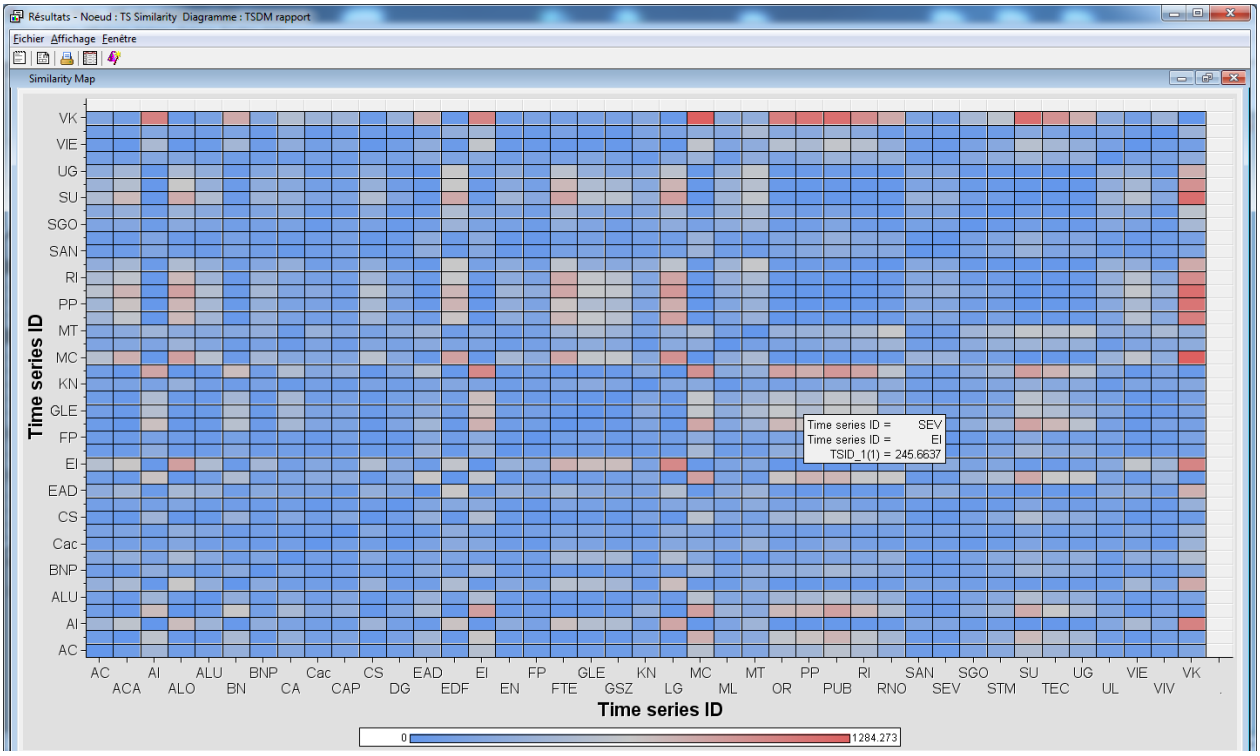
Nous n'allons pas détailler ici le fonctionnement de SAS Enterprise Miner. Nous allons juste expliquer comment obtenir une « similarity map » sur cette solution.
Pour ce faire, il faut créer une source de données contenant la table avec nos 41 séries de cours boursiers.
Pour cela il suffit de cliquer droit sur « Sources de données » à gauche, pour en créer une nouvelle et suivre les instructions à l'écran. Au moment de la sélection de la table, si vous ne la trouvez pas, il vous faut annuler votre manipulation, aller dans le menu « Fichier » → « Nouveau » → « Bibliothèque… » et suivre les instructions (en choisissant comme emplacement de la bibliothèque le dossier contenant votre table).
Lorsque votre bibliothèque contenant la table d'application est créée, recommencez la manipulation en rapport avec la source de données, cela vous permettra de la créer sans difficulté. En plaçant la source sur un diagramme, il vous suffit de rajouter un nœud de préparation de données pour séries temporelles (« TS Data Preparation »). Ensuite, vous ajoutez un nœud de similarité, qu'il restera à paramétrer.



Il est à noter que pour ce genre d'étude, il faut que les différentes observations soient les différentes dates et que les différentes séries soient les variables (appelées classement par Time Series ID en SAS, comme expliqué à la section précédente). Dans notre cas, la table est construite de cette façon, aucune préparation n'est donc nécessaire. En revanche, en fonction du paramétrage de votre source de données, il est possible que vous soyez obligé de passer par l'outil « Métadonnées » pour définir « Date » comme « identifiant temporel continu », et toutes les autres variables en « explicatives continues», à l'exception de « ctrl » qu'il faut « rejeter ».



Ensuite, il ne vous reste plus qu'à exécuter le dernier nœud, qui exécutera automatiquement ceux qui le précèdent, et à cliquer sur « Résultats » quand la fenêtre apparaît.

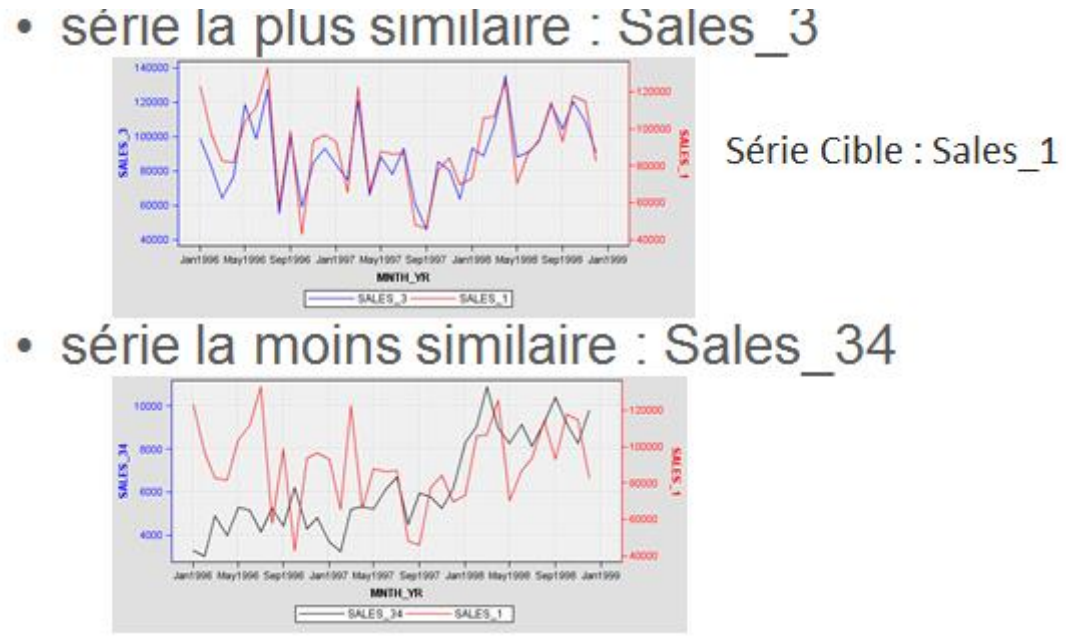


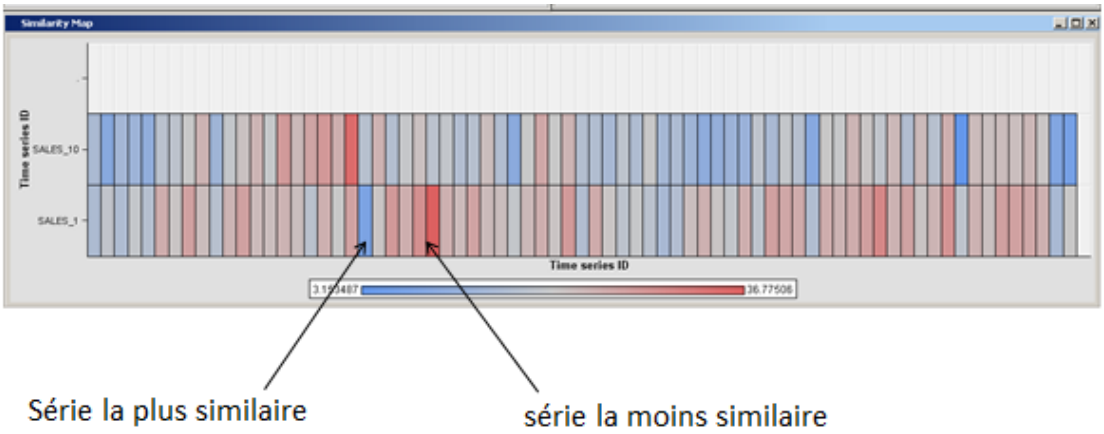
Similarity Map

On voit que la plupart des titres sont assez similaires, à l'exception de quelques-uns très dissimilaires (VK Vallourec et MC, LVMH par exemple). Ce résultat semble logique, il est rare que les titres aient des comportements indépendants les uns des autres. Mais il est possible que certains agissent différemment en fonction de paramètres (secteur, taille,…) qu'il faudra étudier plus précisément pour pouvoir expliquer de manière sérieuse ces différences. On voit que ce résultat nous offre une solide

base de travail et nous permet d'avoir une vue d'ensemble du problème. Il met aussi en évidence le fait de devoir travailler les séries une à une.
Cette petite application nous montre bien la facilité d'utilisation du TSDM et la grande quantité d'information qui en découle et qu'il est possible d'exploiter par la suite, notamment pour faire du clustering hiérarchique.

Si vous souhaitez analyser une (ou plusieurs, cela ne change rien) variable(s) en fonction des autres, il suffit d'avoir cette variable typée « A expliquer » alors que les autres sont explicatives.
On obtient deux graphiques ainsi qu'un extrait de la carte de similarité (suite de l'exemple des cosmétiques) :





3.3. « Time Series Exponential Smoothing »

Le nœud « Time Series Exponential Smoothing » génère des prévisions en utilisant les modèles de lissage exponentiel qui ont optimisé les poids de lissage pour les données des séries temporelles. Le nœud offre des prévisions des modèles de lissage exponentiel simple, double, linéaire, tendance amortie (« damped trend » en anglais), additif saisonnier, multiplicatif saisonnier, multiplicatif de Winters et additif de Winters. Il fournit également des modeleurs ayant la capacité de détecter et de remplacer les outliers, d'exporter des matrices de distances et d'étendre les séries temporelles d'input aux valeurs futures. Vous trouverez dans le SAS/ETS User's Guide, au chapitre 16, par exemple, sur la procédure FORECAST, des informations qui vous décriront en détails ces modèles et les méthodes associées.

Ce nœud peut paraître superflu au vu des nœuds existants mais il est fort utile dans certaines circonstances puisqu'il prévoit facilement et rapidement une grande quantité de séries avec des modèles, certes simples, mais très adaptés à certaines séries.

4. EN CAS DE PROBLEME

4.1. Les problèmes les plus fréquents

Deux problèmes arrivent fréquemment :

- Lors de l'utilisation du nœud « Metadata » la traduction française de « target » (au niveau des rôles des variables) est bien « A expliquer » et non « Prédictions ».
- Si dans le nœud "TS Similarity", on choisit pour l'option "Classification hiérarchique" la valeur "Oui", alors le logiciel ne tient pas compte des variables cibles et n'affiche pas les distances entre les cibles et les inputs (donc pas de matrice de similarité).

4.2. Contacter le Support Clients

Si vous rencontrez des problèmes lors de l'importation ou l'exportation de fichiers dans JMP, vous pouvez nous écrire à support@sas.com, ou remplir notre formulaire en ligne, en précisant le type de fichier en cause, et en joignant, le cas échéant, une copie d'écran de l'erreur reçue, ou encore en décrivant le comportement observé et non attendu.

5. LIENS UTILES

Les nouveautés de SAS Enterprise Miner 12.1
What's New in SAS Enterprise Miner 12.1

Documentation exhaustive sur SAS Enterprise Miner 12.1

Un papier écrit sur le sujet (Mining Transactional and Time Series Data) rédigé par l'un de nos développeurs. Papier intéressant sur ce sujet par un membre de la R&D de SAS:
http://biocomp.cnb.csic.es/~coss/Docencia/ADAM/Sample/Session0_Leonard.pdf
http://www2.sas.com/proceedings/sugi30/080-30.pdf

SAS Introduction pratique : du data management au reporting par Sébastien RINGUEDE

6. CONCLUSION

L'onglet « Séries Chronologiques » qui contient les nœuds vous permettant de faire du Time Series Data Mining, discipline qui connaît actuellement une grande expansion. Cette situation peut dorénavant changer puisque nous sommes dans un contexte où les données sont de plus en plus nombreuses et très souvent sous un format temporel. Ce qui nous incite à tirer parti au maximum de cette information.

SAS a pour l'heure, mis trois nœuds à votre disposition dans SAS Enterprise Miner. L'un pour préparer vos données « Time Series Data Preparation », le deuxième « TS Exponential Smoothing » pour faire des prévisions (lissage exponentiel) sur un nombre très important de données et enfin, le plus important, « TS Similarity » qui sert à identifier la distance et/ou la similarité entre vos différentes séries temporelles.

Cela vous permettra de classer aisément un nouveau produit en fonction de la similarité qu'il a avec une série existante dont le comportement est connu. Vous pourrez aussi faire du clustering et de la réduction de dimension, domaines que nous savons tous pouvant être essentiels.

Jérémy NOEL
Consultant Support Clients SAS France