Projet de thèse : Modèles probabilistes pour l'analyse et la prévision des ventes d'articles dans la grande distribution

Christophe Biernacki et Assi N'Guessan

22 mai 2009

1 Problématique métier et formalisation scientifique

1.1 Problématique de la prévision des ventes dans la grande distribution

1.1.1 Contexte

Pour apporter des solutions adaptées dans le domaine de la grande distribution tant au niveau commercial qu'opérationnel, Soft Solutions travaille depuis plusieurs années à la création de modèles permettant de comprendre les archétypes de ventes d'articles dans le domaine de la distribution. Sa suite ***ibs*** **Retail Business Suite** inclue aujourd'hui un moteur de prévisions des ventes articles en magasin et un moteur d'optimisation des prix de ventes basé sur l'analyse des effets d'élasticité.

Ces solutions sont en grande partie construites à partir de la connaissance du métier de la distribution et atteignent aujourd'hui un pallier nécessitant le développement de modèles et méthodes nouveaux pour optimiser les résultats.

D'autre part, les taux de précision observés se situent autour de 70% par article et par semaine mais sont encore insuffisants pour permettre un pilotage totalement optimisé de l'activité des distributeurs. De surcroît, le contexte économique actuel fait qu'une modélisation aléatoire plus adaptée s'avère nécessaire pour l'analyse et la prévision des ventes d'articles dans la grande distribution. Une telle modélisation des ventes des articles permettrait une meilleure prévision des ventes et se traduirait immédiatement en bénéfices pour les distributeurs.

1.1.2 Objectifs

Dans ce contexte, les objectifs de recherche et les retombées opérationnelles sont :

- améliorer la qualité des modèles actuels et en retirer un gain de précision entre 5% et 10%,
- conserver une approche orientée métier des solutions en proposant des modèles compréhensibles par les experts métiers,
- proposer des solutions évolutives, à même de s'adapter aux changements des comportements des articles ou des consommateurs,
- prendre en compte les contraintes volumétriques inhérentes au sujet (par exemple 50000 articles à modéliser pour 5000 points de vente) afin de construire des solutions directement opérationnelles,
- participer conjointement au rayonnement scientifique et aux diverses publications liées à ce travail.

1.2 Formalisation du problème

Les bases de données (en particulier les volumes de ventes) en grande distribution sont de nature complexe car elles sont mesurées (ou observées) au cours du temps par rapport à des variables (continues, de comptage ou discrètes) associant des facteurs (continus, dicrets ou catégoriels). La modélisation et l'analyse de ces données soulèvent donc des problèmes statistiques et probabilistes importants dus à la non-indépendance des observations répétées dans le temps et à des inter-corrélations entre les facteurs associés.

Ainsi, le volume de vente (y) d'un article est notoirement dépendant d'un certain nombre de facteurs comme le temps (t), le type d'article (a), le magasin ou le lieu de vente (m), le prix en valeur absolue ou en valeur relative par rapport à la concurrence (p), la survenue d'événements particuliers comme une promotion, une publicité, un jour férié notable (e). D'autres facteurs connus ou inconnus, contrôlés ou incontrôlés (x) peuvent aussi s'ajouter à cette liste. Le problème de la prévision des ventes est donc de définir puis d'estimer la forme fonctionnelle suivante :

$$y(t, a, m, p, e, x). \tag{1}$$

Le modèle envisagé devra non seulement être pertinent pour la prévision des ventes (aspect *décisonnel*) mais aussi permettre une description intelligible du mécanisme de vente sous-jacent (aspect *descriptif*). En outre, du fait de la modélisation par nature simplificatrice et de la présence de facteurs incontrôlés et/ou inconnus x, l'incertitude engendrée devra être prise en compte sous forme stochastique.

Dans ce contexte, un *modèle additif* est une approche classique pour décomposer de façon simple et souvent réaliste les effets des différents facteurs et de leurs interactions. Ce modèle pourrait alors servir de modèle de référence dans un premier temps tout au moins. Il s'exprime par :

$$\begin{aligned} y(t, a, m, p, e, x) \quad = \quad & [\text{effets purs}]\ y(t) + y(a) + y(m) + y(p) + y(e) \\ & + [\text{effets d'interactions ordre 2}]\ y(t, a) + y(a, p) + \dots \\ & + [\text{effets d'interactions ordre } >2]\ y(t, a, p) + \dots \\ & + [\text{effet moyen}]\ y + [\text{erreur}]\ y(x). \end{aligned} \tag{2}$$

Le problème se ramène donc à identifier les formes fonctionnelles des « briques de base » que sont les $y(t)$, $y(a)$, $y(t, a)$, ...Typiquement, la brique uniquement temporelle $y(t)$ s'appuiera avec profit sur des modèles probabilistes classiques en séries chronologiques (moyennes mobiles, SARIMA, GARCH). D'autres briques devront faire aussi l'objet d'une étude bibiographique spécifique mais il est prévisible que certaines d'entre elles seront à modéliser « de toutes pièces » car car très proches du contexte métier initial. Ce processus de modélisation sera donc un élément important du sujet de thèse et favorisera largement l'approche probabiliste pour les raisons évoquées précédemment. Il faudra veiller également à ce que chaque modèle de brique de base reste suffisamment signifiant pour être aisément interprétable par des acteurs de la grande distribution non statisticiens de formation.

Les différents modèles proposés pourront alors être validés par des méthodes classiques ou spécifiques de choix de modèles en statistique mathématique. Cela permettra par exemple de sélectionner automatiquement entre deux modèles concurrents sur l'impact d'une publicité dans le temps (brique $y(t, e)$) au fur et à mesure de la disponibilité des retours de vente.

De nouveaux modèles pourront être également proposés en créant des typologies de facteurs. Par exemple, l'effet article est peut être indépendant de la marque de cet article. Dans ce cas, il est préférable d'utiliser un modèle sur la gamme d'article A (modèle $p(A)$)

au lieu d'une modélisation plus fine sur chaque article individuel a (modèle $p(a)$). La fabrication de ces typologies étant délicate et fastidieuse, des méthodes de classification automatique probabiliste pourront être utilement mises en œuvre. Au besoin, des méthodes spécifiques pourront être développées.

1.3 Articulation des aspects méthodologique et opérationnel

Le sujet de thèse se décompose en plusieurs niveaux, permettant de passer du niveau méthologique au niveau opérationnel de l'entreprise partenaire :

1. niveau *modélisation* :
   - formalisation du lien entre les facteurs et les ventes,
   - choix des facteurs influents,
   - sélection de modèles,
   - typologie de facteurs ;
2. niveau *validation métier* :
   - implémentation en laboratoire de la méthode proposée,
   - évaluation de cette méthode sur des données réelles issues de la grande distribution ;
3. niveau *application métier* :
   - aide à l'implantation dans une solution opérationnelle développée par Soft Solutions ;
   - transfert de compétences aux différents acteurs pour les nouveaux modèles proposés ;
   - application aux domaines métier connexes (optimisation des prix de ventes, optimisation des stocks et des approvisionnements, optimisation des assortiments de produits ...)

1.4 Références bibliographiques utilisées

1.4.1 Séries temporelles et modèles linéaires

Anderson T.W. (1994). *The Statistical analysis of times*, John Wiley and Sons.

Brockwell P.J., Davis R.A. (1991). *Time Series : Theory and Methods*, Springer-Verlag.

Galway N. W. (2006). *Introduction to Mixed Modelling*, John Wiley.

Verbeke G., Molenberghs G. (2000). *Linear Mixed Models for longitudinal data*, Springer Verlag, New York.

Zhang J.T., Wu H. (2006). *Nonparametric Regression Methods for longitudinal Data Analysis, Mixed Effects Modeling Approaches*, John Wiley.

1.4.2 Analyse des données, data mining, classification

Agresti A. (2002). *Categorical Data Analysis*, John Wiley.

Duda R.O., Hart P.E., Stork D.G. (2000). *Pattern classification*, 2nd edition, Wiley.

Govaert G. (2003). *Analyse des données*, Hermes.

Lebart L., Piron M., Morineau A. (2000). *Statistique exploratoire multidimensionnelle*, Editions Dunod.

Nakache J.P., Confais J. (2003). *Statistique explicative appliquée*, Editions Technip.

Shmueli G., Patel N.R., Bruce P.C. (2006). *Data Mining for Business Intelligence : Concepts, Techniques, and Applications in Microsoft Excel with XLMiner*, John Wiley.

Tuffery S. (2005). *Data mining et Statistique décisionnelle*, Editions Technip.

1.4.3 Choix de modèles

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 6, 716–723.

Burnham K. P., Anderson D. R. (2002). *Model selection and multimodel inference : a practical information-theoretic approach*, 2nd Edition, Springer-Verlag.

Schwarz G.E. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 2, 461–464.

1.4.4 Prévision de vente en grande distribution

Chang S.H., Fyffe D.E. (1971). *Estimation of Forecast Errors Seasonal-Style-Goods Sales*, Management Sciences, Vol. 18, No. 2, pp. B89 - B96.

Kumar M., Patel N.R., Woo J. (2002). *Clustering Seasonality Patterns in the presence of Errors*, Paper 155, May 2002.

Taylor J.W. (2007) . *Forecasting Daily Supermarket Sales Using Exponentially Weighted Quantile Regression*, European Journal of Operational research, Vol. 178, pp. 154-167.

2 Déroulement et impact de la thèse

2.1 Encadrement du doctorant

La thèse sera co-encadrée scientifiquement par deux chercheurs statisticiens du Laboratoire de Mathématiques, UMR CNRS 8524, Université de Lille 1 :

- Christophe Biernacki, Professeur ;
- Assi N'Guessan, Maître de Conférences HDR.

L'encadrement métier sera assuré, pour le partenaire Soft Solutions, par :

- Sylvain Mongy, Docteur en Informatique et Directeur de l'équipe de recherche ;
- Janine Al-Asswad, Docteur en Chimie et Responsable des relations industrielles.

2.2 Programme et échéancier de travail

La thèse se déroulera sur trois années avec l'échéancier prévisionnel suivant :

- 1ère année :
  - bibliographie et état de l'art sur l'ensemble des thématiques à aborder (5 mois),
  - formalisation du modèle probabiliste (3 mois),
  - implantation d'un algorithme de laboratoire mettant en œuvre cette formalisation sur certains facteurs et interactions prioritaires (2 mois),
  - premières expérimentations sur des données réelles réduites (2 mois).
- 2e année :
  - extension de la méthode à d'autres facteurs et interactions,
  - mise en place d'une stratégie de choix de modèles,
  - classification de facteurs,
  - expérimentation sur des données réelles à plus grande échelle,
  - présentation des résultats à une ou plusieurs conférences internationales.
- 3e année :
  - aide au transfert des méthodes et algorithmes vers une solution opérationnelle,
  - prospection sur les possibilités d'applications connexes,
  - rédaction d'un ou plusieurs articles dans des revues internationales.

2.3 Retombées scientifiques et économiques attendues

Du point de vue scientifique, il s'agira essentiellement de publications scientifiques internationales (revues, conférences).

Du point de vue économique, il s'agira de fournir un avantage concurrentiel certain au partenaire Soft Solutions. Par ricochet, on peut s'attendre à un retour positif sur l'ensemble de la grande distribution.