

ALIEN : Un outil pour modéliser la rédaction de diagnostics économiques

Jean-Louis Roos

Il existe deux façons d'aborder l'analyse d'une série numérique portant sur des données économiques : le statisticien réalisera sur celles-ci des traitements statistiques parmi l'éventail de ceux qui seront adaptés à cette série ; le résultat pourra en être une suite de coefficients, d'indices, de tableaux ou de graphiques. L'économiste, par contre, effectuera un traitement plus "textuel", il commentera les valeurs particulièrement significatives à ses yeux, ou les valeurs récentes, et les recadrera éventuellement dans le contexte économique qu'il connaît.

En fait l'économiste introduira dans son analyse les connaissances qu'il possède sur la signification des chiffres qu'il manipule. C'est-à-dire sur leur sémantique économique.

Un telle dichotomie apparaît évidemment un peu caricaturale : un statisticien-économiste mélangera les deux approches. Pourtant, elles sont fondamentalement différentes : l'analyse statistique relève d'une matière très formelle, peu liée au domaine abordé : on fera les mêmes types de traitement en médecine, en économie ou en météorologie ! ce n'est pas le cas de l'analyse économique qui est spécifique au domaine économique. La différence entre les deux approches est en partie la suivante : en statistique "pure" on introduit peu, ou pas du tout, de sémantique sur le domaine traité. En économie, c'est la sémantique qui prime sur les données.

Si les traitements statistiques sont maintenant bien formalisés, ce n'est pas le cas des discours d'analyse économique. Or, depuis plusieurs années, et un peu partout dans le monde, des chercheurs essaient de trouver une solution au problème suivant : l'être humain, face à une grande quantité d'information sur un sujet donné, est capable – en général – de faire un diagnostic sur ce sujet. Mais comment fait-il ? peut-on formaliser les raisonnements qui conduisent à un tel diagnostic ?

Dans le contexte de l'analyse économique, le problème devient : peut-on décrire, formaliser, modéliser, les mécanismes qui amènent un économiste à "rédiger" tel ou tel diagnostic à partir d'un ensemble de données quantitatives ? Si la réponse est oui, il

serait alors possible d'informatiser ce processus. Or, on vient de le dire plus haut, l'analyse économique littéraire est tout, sauf une science exacte. Donc *a priori* une telle formalisation ne devrait pas être pas possible !

Pourtant, si on observe des commentaires économiques simples, on ne peut qu'être frappé par leur aspect "stéréotypés" – et ce quelle que soit la langue dans laquelle ils ont été écrits – il semble bien qu'il existe des schémas d'analyse très répétitifs, et il devrait être possible de les modéliser. C'est une telle expérience qui a été tentée à l'Insee, d'abord il y a quelques années avec le logiciel Sam, puis actuellement avec le logiciel Alien.

Un tel logiciel fonctionne à ce jour, mais suscite de nombreuses et parfois de vives critiques. On peut classer les opposants à son utilisation en trois groupes :

- ceux qui sont certains que l'analyse économique faite par l'être humain ne peut, en aucune manière, être reproduite par une machine ;
- ceux qui pensent que la reproductibilité serait possible, mais qu'elle est encore trop complexe et trop mal connue. Donc, celle qui peut être faite actuellement restera insuffisante,
- ceux qui s'inquiètent d'un tel logiciel car il ne peut, à terme, que réduire l'intérêt du travail, voire supprimer des emplois.

Bien sûr, tous ces opposants ont partiellement raison. Aucune machine ne pourra égaler une analyse humaine faite par un spécialiste de haut niveau. Par contre, il est vrai qu'il est possible de formaliser des commentaires simples, mais qu'une telle méthode reste complexe et difficile. Enfin le problème des gains de productivité que cela pourrait entraîner reste entier.

Je vais ici aborder uniquement des questions liées aux critiques faites par le second groupe et montrer que s'il y a, certes, des difficultés, beaucoup d'entre elles (mais pas toutes) peuvent être résolues.

En fait Alien est un logiciel fortement paramétré, et si, par défaut, il peut rédiger un texte sur n'importe quelle série économique, il est cependant possible de modifier très fortement la rédaction en fonction des choix que fera l'économiste pour décrire les données qu'il manipule. C'est au niveau de ces choix, on va le voir, que résident les difficultés. Mais Alien n'est pas la cause de celles-ci, il n'en est que le révélateur si l'économiste ne possède pas la totalité de la connaissance initiale qui sera nécessaire.

Avant d'aborder la façon dont Alien construit donc ses jugements, nous ferons une présentation du logiciel.

1. Présentation rapide du logiciel Alien

Alien repose sur l'idée que l'économiste ne manipule pas des séries statistiques, mais plutôt un ensemble de connaissances sur celles-ci, y compris de connaissances sur le vocabulaire associé à chaque série. Alien range la connaissance utile dans des "objets" que l'on nomme les "indicateurs", et qui possèdent toute l'information indispensable à la rédaction d'un diagnostic.

C'est Alien qui construit et renseigne de tels objets. Il effectue ensuite une série de traitements pour arriver à la phase de rédaction.

Ces deux points : la description de l'objet indicateur, et les traitements à réaliser, vont être détaillés ci-dessous. Ils seront complétés par une description du logiciel, et par une présentation des expériences en cours.

1.1 – La représentation de l'information

L'objet que manipule l'économiste est un indicateur, c'est-à-dire une structure où se trouve rangée toute l'information indispensable pour interpréter une (ou plusieurs) série(s). En pratique, une grande partie de l'information étant partageable, un indicateur ne porte que l'information qui lui est spécifique. Le partage de l'information se fait principalement par héritage – ce qui impliquera une organisation hiérarchique des indicateurs – mais aussi par accès à des dictionnaires. Il existe ainsi une information commune à tous les indicateurs.

Le contenu d'un indicateur, et donc l'information qu'il peut porter, est susceptible de modifications d'une application à une autre, ou encore en fonction de développement du système ; les indicateurs sont donc construits à partir de prototypes, c'est-à-dire de modèles, aisément modifiables.

1.1.1 – Un découpage hiérarchique des indicateurs

Une application constitue un sujet qui est décrit dans un indicateur de type sujet. Le sujet possède des domaines. Chaque domaine, décrit dans un indicateur de type domaine, comprend une dizaine d'indicateurs quantitatifs ou qualitatifs dont certains possèdent des occurrences. En gros un domaine correspondra à un paragraphe de texte. La hiérarchie sujet, domaine, indicateur final peut être comparée à celle d'un modèle économétrique où les domaines sont des blocs et les indicateurs sont des équations.

1.1.2 – Les types abstraits

Ce sont des modèles de structure ; ils définissent un objet organisé d'une certaine façon, avec des attributs et des valeurs. Ces types sont abstraits car ils ne représentent jamais un objet réel, ce sont en quelque sorte des modèles ou des prototypes.

Le modèle "indicateur" décrit, sous une forme arborescente, ce que doit contenir un indicateur. Chaque nœud joue un rôle d'attribut. Ces attributs sont en nombre variable. À un attribut peut être associé un sous-arbre ou une valeur. Les valeurs associées dans les indicateurs instances sont aussi modifiables. Voici quelques-uns des attributs existants :

- des types, permettant de classer chaque indicateur. À chaque groupe d'indicateurs sera associé des traitements, et parfois un discours possible particulier ;
- des noms pour l'indicateur ;
- des informations sur la périodicité de la, ou des séries, propre (s) à l'indicateur ;
- des formats d'impression de valeurs ;
- du vocabulaire particulier : verbes, adjectifs, adverbes, prépositions, et des règles grammaticales ;
- des listes d'indicateurs occurrences ;
- des équations ;
- des index de jugement, et des outils permettant de les construire ;
- la description des périodes à analyser en terme de variation et éventuellement d'accélération (ou de décélération) ;
- les valeurs et les noms de la, ou des séries, propre (s) à l'indicateur ;
- des programmes et des traitements.

Cette liste n'est pas exhaustive. Chaque utilisateur peut rajouter de l'information. En pratique, plus d'une centaines d'attributs définissent ainsi la sémantique utile d'un indicateur.

1.1.3 – Les dictionnaires

Ce sont des objets particuliers qui regroupent le vocabulaire commun aux indicateurs, mais aussi la majeure partie des règles grammaticales associées. Le contenu des dictionnaires est défini par un index qui est un type abstrait particulier.

1.1.4 – Les structures narratives stéréotypées

Si les indicateurs sont complètement renseignés, ainsi que les dictionnaires, il est possible d'obtenir une rédaction associée à un diagnostic. Un diagnostic est un ensemble d'informations "codé" dans un vecteur – ou séquence – d'informations que l'on a appelé le "vecteur narratif". Il définit "quoi dire". Ce vecteur est traduit en langue naturelle à travers un vecteur d'ordres de rédaction. Il existe plusieurs vecteurs d'ordres, ou séquences d'ordres, ce sont les "schémas narratifs" ou "structures narratives". Chacun décrit un discours stéréotypé possible, avec ses variantes. Voici des exemples de tels discours stéréotypés relevés dans la presse :

"Sales of new cars in western Europe increased by 1.2 per cent last month, with higher demand in Germany and France compensating for a heavy decline in sales in Italy. Sales in the first 11 months of the year, at 12.53m, were 1.3 per cent lower than in the corresponding period a year earlier." Financial Times Friday December 18 1992

"L'indice composite, qui regroupe onze indicateurs de l'économie américaine, a augmenté de 0.8 % en novembre. Il s'agit de la seconde hausse mensuelle consécutive (+ 0.5 % en octobre) et de la plus forte progression de cet indice depuis le mois de janvier 1992. En novembre ce sont les commandes aux entreprises, les ventes au détail et surtout la confiance des consommateurs qui ont joué." Le Monde 1^{er} Janvier 1993

"Les ventes des détaillants progressent dans l'ensemble en septembre-octobre. Ce mouvement correspond en particulier à une amélioration de l'activité dans les secteurs de l'habillement-textile-cuir et du groupe culture-loisirs-luxe. En revanche les ventes sont stables dans le gros équipement du logement."

Insee Informations Rapides N° 336

1.2 – Organisation des traitements

Pour obtenir la rédaction d'un texte de diagnostic, il est indispensable de fournir au système un minimum d'informations préliminaires ; ne serait-ce que le nom et les valeurs des séries à analyser ! À partir de là, quatre étapes de traitements se succèdent dans ALIEN : la construction de l'arborescence des indicateurs et le remplissage de

ceux-ci, des calculs sur les valeurs, une sélection de l'information pertinente, enfin la rédaction finale. Examinons ces étapes :

1.2.5 – La mécanique d'apprentissage

On comprendra aisément que renseigner un indicateur, avec la totalité de la sémantique qui sera nécessaire à la rédaction, peut être fastidieux (sinon impossible !). Un système d'apprentissage automatique est donc indispensable. Le principe d'une telle mécanique est assez simple : il s'agit d'un moteur qui, à partir d'une information sommaire sur chaque indicateur, et d'une information complète, mais commune à tous les indicateurs, peut construire simultanément les indicateurs et les liens entre ceux-ci ; autrement dit toute l'arborescence du système. Un semblable instrument est cependant forcément insuffisant car l'organisation ainsi construite sera souvent très rudimentaire ; elle ne prendra pas en compte telle ou telle particularité d'un indicateur. Celles-ci devront alors être introduites de façon complémentaire, en se substituant aux données construites automatiquement.

Informatiquement "parlant" l'apprentissage implique l'utilisation de trois fichiers de "connaissances" spécifiques au domaine traité : la description minimale des indicateurs, les informations communes à tous, et des compléments d'informations particulières.

Ceci étant, la phase d'apprentissage implique de nombreux choix a priori qui seront définis par l'économiste : entre autre des choix sur les types d'indicateurs, les périodes analysables, et les paramètres de calculs d'index de jugements.

1.2.6 – L'analyse numérique extensive

Une fois la base de connaissances construite, Alien réalise une analyse extensive des valeurs numériques qui consiste à appliquer sur toutes les périodes de temps analysables – celles-ci ont été définies lors de la phase d'apprentissage – des formules mathématiques prédéfinies. Ces périodes de temps ne correspondent pas à la périodicité de la série, mais aux périodes passées importantes à étudier : par exemple l'évolution à un mois, à trois mois, à un an, etc.

Il s'agit ici d'obtenir le maximum de données quantitatives pour chaque indicateur : les niveaux, mais aussi les variations et parfois les accélérations. Pour chaque information, un jugement qualitatif sera porté et codifié sur une échelle de sept valeurs : de "très faible" à "très fort". Ceci constituera des index de jugement. Ces sept tranches possibles peuvent être réduites, mais non augmentées.

1.2.7 – Les choix narratifs : savoir quoi dire

Après ces deux étapes, Alien dispose d'une masse consistante d'informations. Il est impensable de rédiger un texte à partir de TOUT ce qui a été observé ; un tri est indispensable pour ne conserver que ce qui est pertinent.

Cette étape est essentielle : pour chaque indicateur et domaine, Alien recherche l'information la plus substantielle. Il retient *a priori* l'information la plus "intense", par exemple il sélectionnera les périodes où les jugements portés seront les plus éloignés d'une situation normale. Les choix sur les périodes peuvent aussi se faire à partir de situations "graphiques" : retournement de tendance, accélérations, etc. Ou encore être fixés par l'économiste. Dans un certain nombre de situations, les choix sont faits par le résultat de règles, présentes dans des bases de règles, qui sont des connaissances complémentaires pour Alien.

Finalement le but de cette étape est de définir complètement la sémantique du futur discours d'analyse. Cette sémantique est codée dans le vecteur narratif.

1.2.8 – La rédaction

La dernière étape ne concerne plus que la rédaction d'un texte. Le vecteur narratif est ici traduit dans une langue, par exemple le français – mais une autre langue est possible – à travers des schémas narratifs prédéterminés. L'assemblage du texte se fait à partir de fragments de texte. Alien ne travaille pas sur les éléments de base du langage : noms, articles, adjectifs, verbes,... mais sur des regroupements de ceux-ci : les fragments, qui sont des portions de phrases. Ainsi un fragment peut être un substantif, un adjectif, un groupe nominal ou verbal, ou même encore une proposition complète (groupe nominal et groupe verbal). Les fragments sont associés à une règle d'accord et à un "focus" qui permet de codifier leur signification.

Certains fragments peuvent être décrits directement par l'utilisateur, mais dans la plupart des cas ils sont construits automatiquement à partir d'éléments disponibles dans les dictionnaires, ou dans le "squelette" de l'application. L'assemblage des fragments se fait, après gestion des accords et des conjugaisons, au sein des structures narratives. Celles-ci sont en fait des programmes écrits dans le langage de manipulation des connaissances d'Alien, donc aisément modifiables.

1.3 – Description du logiciel

L'organisation informatique d'Alien comporte deux niveaux : le langage C pour l'écriture d'un interpréteur de manipulation de connaissances, et des bases de connaissances.

1.3.1 – Un interpréteur écrit en langage C

L'interpréteur a pour nom Cia et permet aussi bien un maquettage rapide que le stockage des connaissances. La version 2 actuelle est portée sur plusieurs gammes Unix [SunOs, UP/Ux, Aix, Sco, Coherent], MVS, Dos (uniquement 386 et 486). Le logiciel occupe 150 K et les sources 400K (8000 lignes de langage C). Un manuel est disponible en ligne. L'interpréteur est complété par un ensemble de traitements spécifiques à Alien et aussi écrits en langage C. La rédaction d'une page de texte, apprentissage inclus, prend de dix à cinquante secondes suivant les machines.

1.3.2 – Des fichiers de connaissances

Une application sous Alien repose sur l'utilisation de quelques fichiers spécifiques que l'on doit renseigner *a priori*, ainsi que sur des fichiers, collectifs, en général stables et partagés par toutes les applications.

Les fichiers collectifs.

Ce sont d'une part les dictionnaires, il y en a cinq, plus les bases de règles, et quelques fichiers généraux qui contiennent les prototypes et les structures narratives. Les dictionnaires sont :

- Les synonymes : à chaque entrée correspond six synonymes, les contenus ne sont pas seulement des mots isolés, mais parfois des groupes (nominaux, verbaux), voire des propositions complètes.
- Les conjugaisons : l'entrée est un verbe, ou un groupe verbal, à l'infinitif. Sont conjugués : le présent, le passé, et le conditionnel.
- Les accords : à chaque mot, ou groupe de mots, correspond le même ensemble au masculin singulier, féminin singulier, masculin pluriel, féminin pluriel, neutre singulier, neutre pluriel (ces deux cas sont pour d'autres langues), ainsi que deux cas spécifiques (par exemple en français l'élision avec h).

- Le vocabulaire complémentaire : il est constitué de mots isolés, principalement invariables : prépositions, conjonctions, adverbes, etc.
- Le vocabulaire du manuel : il est possible de disposer pour toute application, d'un manuel complet de celle-ci.
- Les déclinaisons (pour l'Allemand).

Les autres fichiers généraux sont les fichiers d'initialisation, de prototypes et les fichiers de schémas stéréotypés. Ce sont ces derniers qui commandent la rédaction. En principe l'utilisateur ne les modifie jamais (mais il peut le faire). Ils sont liés à la langue utilisée pour la rédaction puisqu'ils commandent la structure des phrases.

Les fichiers propres à l'application

Ce sont ceux que l'utilisateur doit renseigner, "à la main", ou à l'aide de ses propres traitements. On a ainsi :

- les fichiers de la base de données quantitatives : il s'agit des noms et des valeurs séries. On peut aussi associer à ces fichiers un fichier donnant la date pour laquelle on désire un diagnostic.
- le squelette : c'est le descriptif minimal pour CHAQUE futur indicateur : on donnera ici les types de l'indicateur, son nom et la, ou les, série (s) à utiliser,
- le modèle d'indicateur renseigné : c'est un prototype, mais complètement garni avec des valeurs communes pour tous les indicateurs,
- le fichier de complément stylistique : il est facultatif et permet d'améliorer l'apprentissage automatique.

1.4 – Les utilisations d'Alien

1.4.1. – Les enquêtes de conjoncture

Le logiciel est utilisé tous les mois pour l'enquête mensuelle dans l'industrie. Il a été intégré dans la chaîne de dépouillement de l'enquête. Une version permettant de faire des sorties en Dcf/Gml est en cours de réalisation.

Parallèlement des tests sont en cours avec l'enquête Bâtiment et travaux publics (BTP) et ultérieurement auront lieu d'autres expériences avec l'enquête commerce de détail.

1.4.2. - équipements professionnels

29^e projet ICARE

C'est un projet de retour aux entreprises d'un diagnostic conjoncturel sur secteur fin. À titre d'information Alien compose actuellement deux pages de texte pour 41 produits industriels en sept minutes cpu, et demande 12000K pour ce travail. Cependant tous les problèmes techniques ne sont pas encore résolus.

1.4.3. - Le projet IFO

À la demande d'Eurostat et dans le cadre d'un programme de recherche européen (Development of Statistical Expert Systems) l'Insee et l'Institut de recherche économique Ifo à Munich, développeront le produit Alien pour l'adapter à la langue allemande. L'Ifo adaptera le produit à ses besoins, proposera des modifications pour l'améliorer et apportera son expérience des bases de règles dans l'analyse des cycles et dans l'analyse conjoncturelle. Eventuellement avec des propositions sur des méthodes de prévision. Les utilisations que proposera l'Ifo permettront de vérifier si le produit est réellement efficace. Le projet doit durer deux ans.

1.4.4. - D'autres applications

Alien a été prêté, ou est en test, au Central Statistical Office, à la Banque Nationale de Belgique, à la Banque de France et à Eurostat. D'autres expériences sont envisagées.

2. Connaissances, jugements et opinions

Lorsqu'un économiste rédige un diagnostic sur un ensemble de séries, il utilise principalement trois ressources :

- des connaissances particulières sur ces séries : telle série sera un indice, telle autre une quantité absolue, une troisième sera un flux,...L'économiste doit savoir s'il peut, ou non, calculer des variations, et sur quelles périodes ; il connaît, enfin, le vocabulaire, parfois le "jargon", associé à chaque série.
- des connaissances sur la façon dont, pour chaque série, se forment des jugements : l'économiste sait quand il peut parler d'"un fort recul de la production", d'"une hausse des prix vive",...

- des connaissances complémentaires : notre économiste peut expliquer telle ou telle évolution par des informations "exogènes" qu'il possède : par exemple, il interprétera une évolution du prix du baril de pétrole par une certaine situation au Moyen-orient.

On aura noté que ces "trois ressources" apparaissent comme étant, relativement éloignées les unes par rapport aux autres, de moins en moins "objectives" : si on demande une analyse, sur un ensemble de séries, à plusieurs économistes, ils devraient avoir la même position en ce qui concerne les valeurs, peut-être des positions différentes pour ce qui est des jugements à porter, probablement des avis divergents pour ce qui touche aux commentaires et aux opinions.

Dans Alien ces trois ressources doivent être formalisées, et donc programmées. Il existe un paramétrage du système, avec des paramètres par défaut. Cependant il faut prendre conscience que le choix de paramètres différents peut, pour une même série, entraîner la rédaction de diagnostics différents !

2.1 – La connaissance particulière sur chaque série

Nous ne rentrerons pas dans le détail de toutes les connaissances qui sont intégrées dans un indicateur. Ainsi, à chaque série, ou indicateur, peut être associé d'une part un langage, un "dialecte", particulier ; mais aussi une forme, une structure de discours spécifique qui peut dépendre de traitements propres à l'indicateur. D'autres informations, importantes, mais secondaires, ne seront pas abordées ici.

Les tournures et le vocabulaire propres sont introduits dans un fichier "stylistique" qui est utilisé à la phase d'apprentissage (par exemple "les stocks sont légers", "les prix grimpent", "le déficit s'est contracté"...). Les traitements et les structures narratives qui en découlent dépendent par contre de "types", simple codification à faire *a priori*. Ces types jouent un rôle fondamental que nous allons un peu plus détailler. Trois types principaux sont utilisés :

2.1.1. – Le type central

Une série – dans Alien on parle plutôt d'indicateur, c'est à dire d'un ensemble d'informations autour d'une ou plusieurs série (s) – se caractérise par un type central et des types secondaires. Le type central permet de rattacher la série (ou l'indicateur) à une famille parmi quatre possibles :

- type 0 et 1 : les valeurs de la série à analyser sont significatives d'un niveau, d'un stock, d'une quantité absolue, et ce niveau a une signification économique par lui

même. Par exemple : un effectif salarié, un stock de produits finis, un chiffre d'affaires. Dans le type 1 l'étude de la variation de ce niveau ne pourra être faite, car elle n'aurait pas de sens.

- type 2 : les valeurs de la série à analyser sont significatives d'un niveau, mais celui-ci ne veut rien dire ; c'est le cas d'un indice, comme l'indice des prix, par exemple. On ne peut donc le commenter. Par contre, la variation de la série, entre deux périodes, est significative.
- type 3 : les valeurs de la série à analyser sont ici significatives d'une variation. Le niveau est inconnu. La variation, entre deux périodes, de la série implique une accélération (ou un ralentissement, ou encore d'une stabilité de la variation).

Le choix du type implique des commentaires, des analyses, parfois fort différentes. Voici une même série sur la production pour l'ensemble de l'industrie, et par sous secteur, avec les commentaires associés rédigés par Alien. Ceci a été obtenu en ne modifiant que le type central dans le descriptif de l'application !

	Janvier	Février	Mars	Avril	Mai	Juin
La production industrielle	18	17	15	11	10	5
Biens intermédiaires	13	11	11	10	9	4
Équipements professionnels	29	20	18	22	12	18
Branche automobile et transport	8	- 19	- 11	- 17	- 9	- 12
Biens de consommation	23	15	21	12	18	10

type 0

L'activité industrielle, selon les chefs d'entreprise interrogés à l'enquête mensuelle, reste maussade et connaît une nouvelle fois une sensible baisse. Le recul se note dans les biens intermédiaires et dans les biens de consommation ; tandis que la situation est opposée dans les équipements professionnels.

type 1

L'activité industrielle, selon les chefs d'entreprise interrogés à l'enquête mensuelle, reste maussade pour le troisième mois successif.

type 2

Ce mois-ci, l'activité industrielle continue de reculer sensiblement. Le recul se vérifie dans les biens intermédiaires et dans les biens de consommation ; en revanche dans les équipements professionnels l'activité industrielle semble s'améliorer légèrement.

type 3

En juin, la production industrielle a continué d'augmenter faiblement, mais tout au long du trimestre écoulé, plus lentement qu'au premier trimestre. L'augmentation a été observée notamment dans les équipements professionnels ; à l'inverse dans la branche automobile et transport, la production industrielle a reculé légèrement.

Si le type est mal choisi, le commentaire sera tout simplement faux ! mais ceci ne peut être reproché à Alien : un tel commentaire faux serait aussi bien fait par un être humain qui connaîtrait mal son sujet !

2.1.2 – Les formes de variation

Le type que l'on vient de voir n'est pas le seul à jouer un rôle dans Alien, mais c'est le plus important. Il existe aussi un type pour définir quelles variations sont acceptables, ou valides, par exemple : à un mois, à trois mois, etc. Toutes ces variations possibles possèdent une sémantique particulière : en terme de vocabulaire, d'unité des variables, de forme d'équation de calcul, de définition de la période temporelle. Un cas particulier de variation est celui de la variation d'une série qui elle-même est une variation (le type 3 précédent). La signification que l'on donne au résultat dépend alors de la variation de départ et d'arrivée :

- variation de départ positive et variation d'arrivée supérieure ; on dira par exemple : "la croissance s'accélère"
- variation de départ positive et variation d'arrivée positive mais inférieure ; on dira par exemple : "la croissance se ralentit"
- variation de départ négative et variation d'arrivée inférieure ; on dira par exemple : "la chute s'accélère"
- variation de départ négative mais variation d'arrivée négative et supérieure ; on dira par exemple : "la chute se ralentit"
- variation de départ et d'arrivée en sens contraire ; le commentaire se fera dans le contexte de la série.
- les cas de stabilités dans la variation impliquent aussi un commentaire dans le contexte.

On le voit, les règles d'analyses sont plus complexes ici. A ceci s'ajoute le fait qu'il faut déterminer la façon dont on fait l'étude de la variation : moyenne mobile, comparaisons de périodes de tailles différentes,...

2.1.3 – Les occurrences

Une série, un indicateur, peut s'analyser aussi par ses éléments constitutifs : par exemple l'évolution des prix est souvent étudiée, ou expliquée, par l'évolution de ses composants : les prix par grands groupes de produits. Ces composants forment des indicateurs occurrences qui complètent l'indicateur principal, appelé alors "indicateur maître".

En pratique le traitement d'occurrences fait aussi intervenir une codification en type car il existe des situations différentes, qui impliquent des traitements et des discours différents.

Le type occurrence distingue, entre autre, les occurrences instances des occurrences attributs : l'indicateur principal peut avoir des attributs : les stocks seront ainsi une occurrence attribut pour un indicateur principal qui serait un secteur ; mais les occurrences peuvent aussi être des instances : les stocks, dans un secteur, forment une occurrence instance des stocks pour l'ensemble de l'industrie. Dans ces deux cas les discours associés seront différents.

Il existe aussi d'autres formes possibles du type occurrence : l'indicateur principal peut être vide, dans ce cas il n'y a que des occurrences ; il peut aussi y avoir pour un même indicateur principal des sous-groupes d'indicateurs occurrences. Il s'agit toujours de cas particuliers que nous ne présenterons pas ici, mais qu'Alien sait traiter.

2.2 – La construction de jugement

Dans une analyse, un économiste portera généralement un jugement sur les valeurs qu'il observe. Ainsi une augmentation du chômage, des prix, de la production, sera qualifiée par un expert de faible, de moyenne ou de forte ! Comment de tels qualificatifs se forment-ils ?

En fait il n'y a pas de règles absolues. Ainsi entre 1976 et 1978 l'Insee qualifiait ainsi des hausses de prix :

en 1978 : **faible**, avec 7,9 %

en 1976 : **vive**, avec 9,6 %

modérée, avec 7 %

quasi-stable, avec 2,7 %

Chacun aura constaté que la qualification d'un chiffre se fait souvent dans un contexte historique, mais doit-on toujours faire référence au passé ? ou doit-on avoir une référence absolue ? Dans Alien différentes options sont envisageables :

- l'utilisateur peut proposer des tranches de valeurs permettant de fixer directement les qualifications des jugements ;
- les tranches peuvent être calculées automatiquement à partir de l'historique de la série.

Dans ce dernier cas, la "longueur" de la série est importante et influence les résultats.

Si les tranches sont calculées automatiquement, l'utilisateur peut encore fixer certaines options : Alien définit une valeur centrale, correspondant à une situation moyenne. Pour une série du type 0 ou 1 vue plus haut, c'est la moyenne ; pour les autres c'est la variation nulle : 0 si on raisonne en différence, 1 en rapport, 100 en pourcentage. Mais l'utilisateur peut aussi fixer ces valeurs centrales.

Ensuite, les valeurs proches de cette moyenne, feront parties de la tranche moyenne, et plus on s'éloignera et plus on déterminera les tranches permettant de qualifier : assez faible, assez fort – faible, fort – très faible, très fort (ces adjectifs sont bien évidemment dépendants de la série traitée). L'éloignement de la valeur moyenne est aussi paramétrable.

Alien choisit par défaut de découper en 20 tranches égales la quantité comprise entre la valeur centrale et le maximum (ou minimum) de la série, puis de les regrouper pour obtenir les trois tranches inférieures et les trois tranches supérieures correspondants aux qualifications possibles. Tout ceci peut être modifié "à la main" (en fait dans un fichier de paramètres).

Voici quelques résultats pour une série qui est une variation (le solde de réponse à l'enquête mensuelle dans l'industrie). Le texte qui apparaît est généré automatiquement par l'outil de fabrication du manuel d'une application.

On recherche le maximum, le minimum. La valeur moyenne est fixée à 0. La différence entre le maximum et 0 est divisée par 20. Cette quantité forme un fractile du côté maximum. On agit de même pour définir un fractile du côté minimum. Ensuite trois limites sont calculées automatiquement de part et d'autre de la moyenne :

proche de zero : zéro + 2 fractile (s) minimum ou maximum.
médian : moyenne + 11 fractile (s) minimum ou maximum.
éloigné : moyenne + 18 fractile (s) minimum ou maximum.

Ceci défini donc automatiquement 6 bornes :

- 30.6 ; - 15.3 ; - 3.4 ; 2.6 ; 11.7 ; 23.4

sur les valeurs de variations suivantes :

3 - 1 3 0 - 3 2 3 4 5 6 3 - 2 2 4 4 7 7 2 10 11 12 8 14 13 20 24 18 21 24 20 19 19 20
20 23 25 25 24 22 25 22 26 21 20 20 20 18 21 18 17 15 11 10 5 10 8 5 4 8 - 8 - 12
- 14 - 13 - 17 - 16 - 16 - 12 - 11 - 10 - 11 - 9 - 12 - 21 - 6 - 2 - 6 - 4 - 5 - 6 - 10
- 14 - 21 - 19 - 24 - 27 - 28 - 31 - 34 - 33 - 30 - 42

Ce qui donne les résultats suivants :

Entre - 34. et - 30.6 : La production industrielle a diminué très fortement.
Entre - 30.6 et - 15.3 : La production industrielle a diminué sensiblement.
Entre - 15.3 et - 3.4 : La production industrielle a diminué faiblement.
Entre - 3.4 et 2.6 : La production industrielle paraît se stabiliser.
Entre 2.6 et 11.7 : La production industrielle a augmenté faiblement.
Entre 11.7 et 23.4 : La production industrielle a augmenté sensiblement.
Entre 23.4 et 26. : La production industrielle a augmenté très fortement.

2.3 - L'opinion

A l'analyse et au jugement relativement objectif que peut faire un économiste sur un indicateur, s'ajoute fréquemment une "opinion" qui lui est propre. Voici un exemple choisi dans la presse

"Consumer optimism about the economy rose slightly in August from July. But the gain was small : from 77 in July to 77.3 in August and economist said it did not represent the beginning of a trend."

NEW YORK TIMES, LE 28 AOUT 1993

Ce court diagnostic est bien représentatif de ce qui est fait habituellement. Il y a trois analyses successives : statistique d'abord : un indicateur d'opinions des consommateurs, qui passe de 77 à 77.3. Le texte stéréotypé aurait pu s'arrêter là. Il était alors parfaitement neutre. Un jugement économique complète cependant cette analyse statistique : il s'agit d'une augmentation faible, de peu d'importance. On voit qu'a joué

ici un mécanisme de jugement. Enfin s'ajoute une opinion d'économiste : *on ne peut juger qu'il s'agit du début d'un trend "hausssier"*.

Si les deux premiers types d'analyses peuvent aisément être automatisées dans Alien, ce n'est pas le cas de la troisième : l'opinion fait appel à trop d'informations extérieures. Par contre s'il s'agit de commenter des liens entre indicateurs, ou d'ajouter des "opinions" assez simples et "répétitives" il est possible d'utiliser des bases de règles qui se déclenchent en fonction d'un ensemble d'informations prédéfinies vérifiées. Ces bases sont bien sûr à construire spécifiquement à l'application.

3. Conclusion

Le logiciel Alien n'est pas parfait, loin de là. Tel quel, pourtant, il apporte plusieurs services à l'économiste-statisticien qui peut ainsi compléter utilement des analyses purement statistiques :

3.1 – Le rôle d'un assistant

Lorsqu'une application Alien est intégrée dans une chaîne de dépouillement ou de traitement de données, elle fournit un "plus" à l'économiste : non seulement il a ses tableaux habituels, ses indices, ses coefficients, mais il a aussi un texte tel celui qu'aurait pu lui rédiger un assistant. Ce texte peut être construit directement dans un fichier qu'il pourra ensuite modifier à travers un traitement de texte. L'intérêt de ce premier texte tient dans le fait qu'Alien n'aura rien oublié d'important, mais surtout que ses analyses sont "stables" dans le temps, et totalement non influencées par des facteurs extérieurs.

3.2 – Le commentaire de données trop nombreuses

Il peut arriver d'avoir à analyser un trop grand nombre de situations : dans une analyse de secteurs fin par exemple, ou dans des champs géographiques détaillés. Dans de tels cas, l'être humain ne peut faire face à la rédaction de dizaines, voire de centaines de diagnostics. Alien peut écrire de tels textes, en grande quantité, et rapidement. Il peut aussi commenter automatiquement des tableaux de bords réguliers en faisant ressortir l'information importante.

Signalons enfin que, si dans sa forme actuelle Alien traite de séries chronologiques, toutes autres formes de données économiques peuvent être analysées. Moyennant bien sûr un complément d'analyse et de programmation.

3.3 – La définition de références

Les américains appellent guideline ("références" ?) la définition, dans des domaines divers, de recommandations de "bonnes pratiques". Une bonne pratique est ce qu'il faut faire, dire, écrire face à une situation donnée. Ces guidelines ont été surtout développés en médecine, à l'origine à partir de nombreux travaux de formalisation de diagnostics ou de traitements pour des logiciels d'Intelligence Artificielle. En gros un guideline permet de décrire un consensus formalisé par de nombreux praticiens. Ce consensus, certes très relatif, permet de fournir une réponse efficace et argumentée devant une situation donnée.

Alien permet, dans le domaine économique, de construire de telles références. C'est le cas à travers le projet de manuel automatique de descriptions de l'interprétation d'une application. Ce manuel, construit automatiquement, devrait permettre, pour chaque application dans Alien, de décrire celle-ci et ses règles d'interprétation et d'analyse.

3.4 – Statistique et sémantique des données

Le statisticien et l'économiste manipulent les mêmes chiffres, mais l'économiste a besoin de beaucoup plus de sémantique que le statisticien. Alien est un outil qui permet la constitution et l'utilisation d'une grande partie de la sémantique nécessaire au diagnostic économique. Mais il reste un outil insuffisant, le logiciel est tout petit : à titre d'information la totalité du produit représente 600 K sur une disquette. Des produits plus importants seront sûrement construits dans les années à venir, mais ils demanderont de plus en plus aux statisticiens et aux économistes de s'intéresser à la formalisation, et à la gestion, de la sémantique des données statistiques qu'ils manipulent. Ce sera sûrement un des plus gros défis auxquels ils seront confrontés dans les années à venir.

BIBLIOGRAPHIE

J.L.ROOS, Intelligence Artificielle en Langage C, Editions Eyrolles, (1992).

J.L.ROOS, Un assistant logique pour l'interprétation experte de données numériques, INSEE, Séminaire Recherche du 8 avril 1993.

PS : Une disquette de démonstration sous Dos est disponible (pour Pc 386 ou 486).