Direction des Études et Synthèses Économiques

G 2006 / 12

Théorie de l'opinion
Faut-il pondérer les réponses individuelles ?

Olivier BIAU et Nicolas FERRARI

Document de travail

INSEE

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail
de la Direction des Etudes et Synthèses Économiques

G 2006 / 12

Théorie de l'opinion
Faut-il pondérer les réponses individuelles ?

Olivier BIAU et Nicolas FERRARI **

NOVEMBRE 2006

Les auteurs remercient Matthieu CORNEC pour sa discussion d'une version antérieure de cette étude lors du séminaire du D3E du 26 juin 2006 ainsi qu'Eric DUBOIS, Hélène ERKEL-ROUSSE et Philippe SCHERRER pour leur relecture attentive et leurs précieux conseils.

---

* au moment de cette étude, Olivier BIAU et Nicolas FERRARI travaillaient à la Division des enquêtes de conjoncture -Département de la Conjoncture, Timbre G120 - 15, bd G. Péri - BP 100 - 92 444 Malakoff Cedex.

Département des Etudes Economiques d'Ensemble - Timbre G201 - 15, bd Gabriel Péri - BP 100 - 92244 MALAKOFF CEDEX - France - Tél. : 33 (1) 41 17 60 68 - Fax : 33 (1) 41 17 60 45 - CEDEX - E-mail : d3e-dg@insee.fr - Site Web INSEE : http://www.insee.fr

Théorie de l'opinion
Faut-il pondérer les réponses individuelles ?

Résumé

Les enquêtes de conjoncture fournissent une information avancée et facilement mobilisable par les prévisionnistes. Leurs questions sont le plus souvent qualitatives. L'utilisation la plus courante chez les conjoncturistes consiste à calculer des soldes d'opinion, différences entre les proportions des réponses positives et des réponses négatives. Ces soldes sont ensuite introduits dans des modèles économétriques afin de prévoir les évolutions des variables étudiées.

Les soldes d'opinion sont en général calculés en pondérant les réponses des entreprises. Le plus souvent, les poids sont proportionnels à la taille des entreprises. Toutefois, il n'existe pas d'éléments théoriques justifiant de manière précise de telles pondérations. Nous développons ici un modèle en vue de déterminer le choix optimal des pondérations. Dans l'objectif de construire des soldes d'opinion qui prévoient au mieux les évolutions réelles de la variable sous-jacente, nous mettons en évidence qu'il faut que les pondérations croissent moins que linéairement avec la taille des entreprises. Cette conclusion est vérifiée empiriquement à l'aide de plusieurs exemples de questions issues d'enquêtes de conjoncture menées par l'Insee.

Mots-clés : Enquêtes de conjoncture, quantification, solde d'opinion, prévision conjoncturelle

---

Balance of opinion
What about missing the weights?

Abstract

Due to their early release, Business Tendency Surveys (BTS) are widely used in short term forecasting. Their questions are mainly qualitative; answers are most often used to calculate balances of opinions, which are defined as the difference between the proportions of positive answers with respect to the negative ones. These indicators are then used by forecasters as explanatory variables in econometric models.

The balances of opinions are generally weighted with the firm size. However, there is no theoretical evidence of the efficiency of this kind of weighting. We propose here a model which aims at determining optimum weights; these weights should allow us to optimize the forecast of the macroeconomic variable. According to our analysis, the weights have to grow less than proportionally with the firm size. This conclusion is empirically tested through several examples derived from the French Industry BTS.

Keywords: Business Tendency Surveys, quantification, balance of opinion, short-term forecasting

Classification JEL : C8, C42, C53, E27

Table des matières

1 Introduction

Les enquêtes de conjoncture sont des enquêtes légères, à périodicité infra-annuelle. La plupart des questions posées à ces enquêtes sont qualitatives et appellent une réponse à trois modalités (« en hausse », « stable » ou « en baisse »). Du fait de leur publication très rapide, les résultats des enquêtes de conjoncture permettent en particulier d'effectuer des prévisions à court terme des principaux agrégats des comptes trimestriels (production, emploi...).

L'indicateur le plus couramment utilisé par les conjoncturistes pour résumer les réponses à une question d'une enquête est le solde d'opinion, défini comme la différence entre le pourcentage[1] de réponses positives et le pourcentage de réponses négatives à cette question. Dès l'origine des enquêtes de conjoncture, le débat sur la méthode à utiliser pour agréger les réponses individuelles a été vif. La littérature sur ce sujet permet de distinguer trois approches méthodologiques.

La première, dite « méthode par régression », consiste à construire dans un premier temps des indicateurs simples à partir des proportions des réponses aux trois modalités (« en hausse », « stable » et « en baisse »). Dans un deuxième temps, ces indicateurs sont utilisés dans des modèles de régressions afin de prévoir les évolutions conjoncturelles. De telles méthodes apparaissent dès le début des années 1950. Elles ont été appliquées en premier lieu à l'enquête « the Munich Business Test », conduite à partir de 1950 en Allemagne de l'Ouest par l'IFO. Le solde d'opinion est proposé par Anderson dès 1951. Theil et Cramer (1954) ajoutent un indicateur supplémentaire : le « coefficient de disconformité ». Ce coefficient mesure la dispersion des réponses autour de leur tendance.[2] Plus récemment, Hild (2002) propose de préférer au solde d'opinion une combinaison linéaire des trois proportions de réponses, obtenue par une méthode d'analyse en composantes principales. Au final, le solde d'opinion garde néanmoins un quasi-monopole chez les conjoncturistes.

[1] Ce pourcentage est généralement doublement pondéré - cf. infra.

[2] Si on note R^+ et R^- les proportions de réponses « en hausse » et « en baisse »ainsi que S le solde d'opinion, alors le coefficient de disconformité est donnée par $R^+ + R^- - S^2$.

Pas sa simplicité de mise en oeuvre, il constitue l'indicateur phare recommandé par l'OCDE dans son guide « Business Tendency Surveys : A Handbook ». Même s'il ne résume pas l'ensemble de l'information contenue dans les trois proportions de réponses, le solde d'opinion apporte une information pertinente, mesurée par exemple par sa bonne corrélation avec la variable d'intérêt. Theil (1961) et Fansten (1976) ont cherché des justifications théoriques à l'utilisation des soldes d'opinion à travers un modèle mathématique reliant le solde d'opinion à la variable que l'on cherche à prévoir. Les conjoncturistes acceptent alors comme acquis l'efficacité de cet indicateur et portent d'avantage leur intérêt en aval, avec des méthodes permettant de résumer l'information apportée par plusieurs soldes d'opinion (indicateurs synthétiques[3], indicateurs de retournement[4]).

La deuxième approche, dite « approche probabiliste », trouve son origine dans les prévisions de l'inflation. Dans les années 1970, les conjoncturistes se sont demandés comment utiliser les réponses qualitatives relatives aux évolutions de prix afin de réaliser des prévisions quantitatives de l'inflation. Carlson et Parkin (1975) proposent un modèle probabiliste pour estimer directement la prévision d'inflation moyenne des entrepreneurs à partir de leurs réponses qualitatives. Néanmoins, ces travaux restent relativement théoriques et cette seconde approche est finalement peu utilisée en pratique.

Les débats autour d'indicateurs alternatifs aux soldes d'opinion semblent ensuite faire une pause du milieu des années 1980 jusqu'à la fin de la décennie de 1990.

Une troisième approche s'est faite jour plus récemment et propose des méthodes de construction d'indicateurs tentant de tenir compte de l'hétérogénéité des réponses individuelles aux enquêtes de conjoncture. Sur la construction d'indicateurs alternatifs (« méthode désagrégée »), on pourra se reportera à Mitchell, Smith et Weale (2002). Biau, Erkel-Rousse et Ferrari (2005) appliquent cette méthodologie aux

[3] Par exemple Doz et Lenglart (1999).
[4] Par exemple Grégoir et Lenglart (1998).

données françaises et soulignent certaines difficultés de mise en oeuvre.

Cet article se place dans la première approche en se concentrant sur les soldes d'opinion. En lien avec la théorie des sondages, le solde d'opinion (de la population entière) est estimé à partir des réponses individuelles des entreprises de l'échantillon, agrégées à l'aide de poids individuels, dits « pondérations primaires » et des coefficients de redressement par strate, dits « pondérations secondaires ». La pratique française consiste à utiliser des poids proportionnels à la taille de l'entreprise afin de refléter la structure de l'économie. Toutefois, une comparaison internationale montre que cela n'est pas universel. Par exemple, en Allemagne, l'IFO utilise des pondérations proportionnelles au logarithme des effectifs des entreprises ou encore des pondérations entières déterminées à partir de classes de chiffres d'affaires[5]. Fansten (1976) précise que pondérer ou non les réponses des entreprises ne remet pas en cause le bien-fondé de l'utilisation du solde d'opinion ; l'OCDE souligne dans son handbook que le calcul des soldes est robuste au choix des pondérations primaires[6]. Il nous semble toutefois que les séries de soldes d'opinion peuvent être plus ou moins prédictives selon les pondérations individuelles choisies.

Cet article a pour objectif d'éclairer de manière théorique ce choix. Il apparaît que les pondérations optimales sont bien croissantes avec la taille de l'entreprise, mais que cette croissance doit être moins que proportionnelle. Ce résultat est confirmé par quelques exemples empiriques.

L'intuition de ce résultat est relativement simple. Les enquêtes de conjoncture sont des enquêtes avec des échantillons de sondage qui ne couvrent qu'une partie des entreprises. Il faut donc considérer l'agrégat à prévoir comme la somme de deux

[5]Par exemple : 1 si les chiffres d'affaires est inférieur à 0,25 millions d'euros, 2 si le chiffre d'affaires est compris entre 0,25 et 0,5 millions d'euros, etc...

[6]« Furthermore, practical experience has shown that the balances are not very sensitive to the choice of weighting variables. In practice it is sufficient to use a single variable reflecting the general economic importance of the enterprise in weighting all the survey answers. »

composantes : celle des entreprises de l'échantillon (partie *in sample*) et celle des entreprises extérieures à l'échantillon[7] (partie *out of sample*). Les évolutions de l'agrégat *in sample* résultent directement des comportements des entreprises couvertes par l'enquête. Il y a donc un lien d'ordre comptable entre ces évolutions et les réponses des entreprises. Des pondérations proportionnelles sont donc pertinentes pour la prévision de la composante *in sample*. Au contraire, seule une composante conjoncturelle commune lie les réponses des entreprises de l'échantillon aux évolutions de la composante *out of sample*. Dans ce cas, à l'instar de ce qui est fait actuellement dans les enquêtes de conjoncture pour les questions sur les perspectives générales, sous couvert que les strates sont suffisamment homogènes, il n'y a pas lieu de pondérer les réponses des entreprises[8] pour la prévision de l'agrégat *out of sample*. Au final, les pondérations optimales doivent donc bien croître avec la taille de l'entreprise, mais non proportionnellement.

Le modèle théorique proposé dans cet article est inspiré de Fansten (1976). Toutefois, le modèle initial est fortement amendé afin de porter un regard attentif sur la question des pondérations primaires. Ce travail permet en outre de soulever de nouvelles réflexions théoriques relatives à l'introduction des soldes d'opinion. En particulier, ne garder que la différence entre les proportions de « hausse » et de « baisse » repose sur des hypothèses assez fortes qui ne sont pas nécessairement vérifiées en pratique.

L'utilisation des réponses à des questions qualitatives pour la prévision d'évolution d'agrégats économiques est introduite par la partie 2. Le développement théorique se décompose ensuite en deux étapes, correspondant aux parties 3 et 4. La première étape repose sur une hypothèse simplificatrice qui permet d'alléger les calculs et de donner l'intuition du résultat. Par exemple, concernant des questions

[7] Nous entendons en réalité par l'échantillon uniquement les entreprise répondant à l'enquête et non pas l'ensemble des entreprises interrogées.

[8] En effet, l'opinion sur les perspectives générales de production pour l'ensemble de l'industrie française formulée par un chef d'entreprise, c'est à dire sa perception de la conjoncture industrielle, est a priori indépendante de la taille de l'entreprise qu'il dirige.

sur la production, on suppose que les réponses des autres entreprises n'apportent pas d'information supplémentaires aux réponses de l'entreprise dont on veut prévoir la production. Dans ce cas, les pondérations optimales pour les prévisions de l'agrégat *in sample* seraient alors des pondérations proportionnelles à la taille des entreprises (partie 3.1). Au contraire, pour les prévisions de l'agrégat *out of sample*, il ne faut pas pondérer les réponses individuelles (partie 3.2). Finalement, les pondérations optimales seraient des pondérations intermédiaires entre des pondérations constantes et des pondérations proportionnelles (partie 3.3). Dans le cas extrême d'un taux de couverture[9] de 100%, il faudrait choisir des pondérations proportionnelles. La deuxième étape (partie 4) lève l'hypothèse simplificatrice de la partie 3. Des pondérations intermédiaires sont alors préfèrables pour la prévision de l'agrégat *in sample* (partie 4.1). Ainsi, même dans le cas d'un taux de couverture de 100%, les pondérations primaires devraient croître moins que proportionnellement avec la taille des entreprises (partie 4.2). On envisage alors la mise en place de manière pratique de telles pondérations (partie 4.3). La partie 5 vérifie ces résultats théoriques pour la prévision de trois grandeurs économiques : les effectifs salariés de l'industrie manufacturière (partie 5.1), la production manufacturière (également partie 5.1) et les investissements industriels (partie 5.2). Dans chaque cas, nous utilisons des soldes d'opinion issus des enquêtes de conjoncture dans l'industrie de l'Insee (enquête sur l'activité dans l'industrie et enquête sur les investissements dans l'industrie). Enfin, la partie 6 conclut sur l'intérêt d'envisager des pondérations non-proportionnelles à la taille des entreprises.

2 Utilisation des enquêtes qualitatives en prévision conjoncturelle

Dans le cadre de la prévision conjoncturelle, il est habituel de prévoir des variations comptables (comptes nationaux trimestriels, indice de production industrielle, ...) à l'aide d'indicateurs tirés d'enquêtes. Pour cela, on utilise des modèles d'étalonnage

[9] Le taux de couverture est défini ici comme le ratio du total des chiffres d'affaires des entreprises de l'échantillon rapporté à celui des entreprises de la population.

univariés (régressions linéaires) ou multivariés (modèles VAR). De manière générale, tous ces modèles sont linéaires. Il est donc naturel de rechercher des soldes d'opinion qui soient le mieux corrélés possible aux variations comptables étudiées. On veut donc que les soldes d'opinion construits soient proches, à une transformation affine près, de ces variations.

Les entrepreneurs sont interrogés dans les enquêtes de conjoncture sur les évolutions passées et prévues de certaines quantités économiques individuelles. Ces questions sont qualitatives, avec en général trois modalités : « en hausse », « stable » et « en baisse ». Par exemple, pour fixer les esprits, les industriels indiquent ainsi leurs prévisions sur l'évolution de leur production propre durant les trois prochains mois. Dans ce cas, leurs réponses sont naturellement rapprochées des évolutions prévues de la production manufacturière pour le trimestre à venir.

Les réponses sont agrégées sous forme de solde d'opinion, différence des proportions de réponses « en hausse » et des réponses « en baisse ».

Plus précisément, les échantillons des enquêtes de conjoncture sont stratifiés. Les soldes d'opinion sont alors calculés en deux temps comme indiqué ci-dessous. On considère une occurrence de l'enquête, telle que la question d'intérêt porte sur le trimestre t. Par exemple, concernant la question relative à la production prévue dans l'enquête sur l'activité dans l'industrie de l'Insee, pour l'enquête de janvier de l'année A, t est le premier trimestre de l'année A. On note s_i la réponse de l'entreprise i en t. s_i est une variable aléatoire qui est à valeur dans $\{-1, 0, 1\}$. Ainsi, lors de cette enquête :

1. On calcule tout d'abord au sein de chaque strate h les parts R_h^s de réponse par modalités s. Ces parts sont des pourcentages pondérés avec les poids individuels $\{p_i\}_i$. Les poids individuels sont construits de manière à ce que leurs sommes par strate soient égales à 1 : $\sum_{i \in h} p_i = 1$.

$$\begin{aligned} R_h^+ &= \sum_{i\in h} p_i \mathbf{1}_{\{s_i=\ 1\}} \\ R_h^0 &= \sum_{i\in h} p_i \mathbf{1}_{\{s_i=\ 0\}} \\ R_h^- &= \sum_{i\in h} p_i \mathbf{1}_{\{s_i=-1\}} \end{aligned}$$

2. Dans un deuxième temps, les proportions pondérées par strate sont agrégées à l'aide de coefficients de redressement (P_h pour la strate h). Ces coefficients reflètent la structure de l'économie et, là encore, ils sont construits de manière à ce que leur somme soit égale à 1 : $\sum_h P_h = 1$. Pour la modalité s, la proportion de réponses à cette modalité est notée R^s.

$$\begin{aligned} R^+ &= \sum_h P_h R_h^+ \\ R^0 &= \sum_h P_h R_h^0 \\ R^- &= \sum_h P_h R_h^- \end{aligned}$$

Le solde S est calculé en faisant la différence entre les proportions « en hausse » et les proportions « en baisse ».

$$S = R^+ - R^-$$

En outre, dans chaque strate h, on définit le solde d'opinion par $S_h = R_h^+ - R_h^-$. On a alors naturellement :

$$S = \sum_h P_h S_h$$

Les pondérations individuelles intra-strates p_i sont appelées pondérations primaires. Elles peuvent être constantes (cas « non pondéré ») ou fonction d'une variable individuelle mesurant la taille de l'entreprise. Traditionnellement les pondérations sont proportionnelles à la grandeur économique dont on veut estimer les variations. Dans la suite, nous appellerons cette grandeur « variable d'intérêt » et ses variations « variations d'intérêt ». Par exemple, pour étudier les évolutions prévues de l'emploi salarié, les pondérations utilisées seront les effectifs salariés des entreprises (rapportés à la somme de ces mêmes effectifs dans la strate). Souvent la

variable d'intérêt n'est pas mesurée par l'enquête et on choisit alors une variable qui, en première approximation, lui est fortement corrélée. Ainsi, concernant les évolutions de la production, la variable choisie comme pondération sera le chiffre d'affaires de l'entreprise.

Les pondérations primaires sont fixées de manière très empirique. En particulier, des pondérations proportionnelles à la taille de l'entreprises sont souvent justifiées par le souci de refléter au mieux les variations de la variable d'intérêt, sans pour autant s'appuyer sur une modélisation précise. Dans son article de référence sur la théorie de l'opinion, Fansten (1976) montre que les soldes apportent de l'information pour la prévision de la variable d'intérêt, ceci quelles que soient les pondérations choisies. L'objet de cet article est de proposer un cadre de réflexion pour le choix du type de pondérations (pondérations constantes, proportionnelles à la taille, ...).

Nous ne traiterons pas ici des pondérations secondaires. Aussi, pour simplifier, le développement théorique écrit ci-après portera sur une unique strate afin de simplifier le modèle. L'objectif est de prévoir au mieux les variations de la variable d'intérêt dans la population des entreprises qui correspondent à une strate unique de l'enquête. Les strates sont déterminées de manière à être les plus homogènes possible. Classiquement, elles seront définies par les secteurs ou les branches d'activité et par un critère de taille.

Considérons que l'évolution à prévoir est celle de la variable agrégée X. Par exemple, X peut représenter la production manufacturière pour le trimestre t considéré. Soit T le taux de croissance associé. On note x_{-1} la valeur de la variable X à la période précédente. x_{-1} est considéré connu et non aléatoire.

$$T = \frac{X}{x_{-1}} - 1$$

En notant $\mathcal{P}$ la population toute entière de la strate et T_i les évolutions de la variable

individuelle X_i pour l'entreprise i :

$$\begin{cases} X = \sum_{i\in\mathcal{P}} X_i \\ x_{-1} = \sum_{i\in\mathcal{P}} x_{i,-1} \\ T = \sum_{i\in\mathcal{P}} q_i T_i \text{ avec } q_i = \frac{x_{i,-1}}{\sum_{j\in\mathcal{P}} x_{j,-1}} \end{cases}$$

Il est alors possible de distinguer la part des variations due à l'échantillon $\mathcal{E}$ des répondants et la part due aux entreprises de la population $\mathcal{P}$ qui n'appartiennent pas à l'échantillon $\mathcal{E}$ des répondants. Ce deuxième ensemble d'entreprises, hors échantillon $\mathcal{E}$, est noté $\mathcal{P} \setminus \mathcal{E}$. En notant respectivement $T^{\mathcal{E}}$ et $T^{\mathcal{P}\setminus\mathcal{E}}$ les évolutions de la variable d'intérêt dans $\mathcal{E}$ et dans $\mathcal{P} \setminus \mathcal{E}$, on peut écrire :

$$T = Q\,T^{\mathcal{E}} + (1-Q)\,T^{\mathcal{P}\setminus\mathcal{E}} \text{ avec } Q = \frac{\sum_{i\in\mathcal{E}} x_{i,-1}}{\sum_{i\in\mathcal{P}} x_{i,-1}}$$

Le solde d'opinion, résumé de l'information donnée par les entrepreneurs, est utilisé afin de prévoir les évolutions agrégées T mesurées *a posteriori* par la comptabilité nationale trimestrielle. Le solde d'opinion S est donc susceptible d'apporter une information relativement à $T^{\mathcal{E}}$ et à $T^{\mathcal{P}\setminus\mathcal{E}}$, mais ceci selon deux modalités différentes :

– Les réponses individuelles renseignent directement - de manière comptable - sur les évolutions *in sample*, c'est-à-dire sur les évolutions $T^{\mathcal{E}}$ sur l'échantillon $\mathcal{E}$.

– Concernant les évolutions *out of sample*, c'est-à-dire les évolutions de $T^{\mathcal{P}\setminus\mathcal{E}}$ (hors de l'échantillon), il n'y a pas de lien comptable entre les évolutions individuelles mesurées par les réponses à l'enquête et les évolutions de l'agrégat. Le lien est « économique » et réside dans une situation conjoncturelle globale qui touche l'ensemble de la population des entreprises de la strate, qu'elles répondent ou non à l'enquête.

3 Choisir les pondérations primaires, modèle simplifié

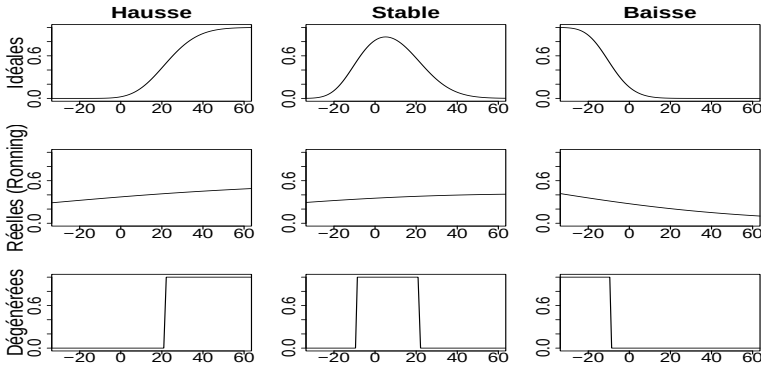
3.1 Pondérations optimales pour l'estimation *in sample*, modèle simplifié

Supposons que les évolutions T_i sur l'échantillon $\mathcal{E}$ suivent une loi aléatoire, identique au cours du temps, de moyenne m_T et de variance σ_T^2. Soit f_T et F_T sa densité et sa fonction de répartition. On appelle $\mathcal{I}_{-1}$ l'information disponible à la fin de la période $t-1$. $\mathcal{I}_{-1}$ contient en particulier les grandeurs qui permettent de calculer q_i, le poids accordé en t à l'entreprise i dans l'agrégation des réponses. En t, les soldes sont connus et l'évolution courante T_i est à prévoir. Par linéarité, $E\left[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_j\}_{j\in\mathcal{E}}\right]$ est donné par $\sum_{i\in\mathcal{E}}\frac{q_i}{Q}E\left[T_i|\{s_j\}_{j\in\mathcal{E}}\right]$.

Supposons dans cette partie que les réponses des autres entreprises de l'échantillon n'apportent pas d'information supplémentaire à la réponse de l'entreprise i pour prévoir l'évolution individuelle T_i. Cette hypothèse peut être assez réaliste dans la mesure où les déterminants conjoncturels - communs à toute la strate - sont faibles par rapport aux déterminants individuels. De ce fait, $E\left[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_j\}_{i\in\mathcal{E}}\right]$ devient alors égal à $\sum_{i\in\mathcal{E}}\frac{q_i}{Q}E\left[T_i|s_i\right]$.

A une évolution T_i donnée, toute entreprise ne donnera pas nécessairement la même réponse. Aussi, la réponse d'une entreprise i est une fonction aléatoire de l'évolution T_i mesurée (ou anticipée) par l'entreprise. Theil (1961) appelle « fonctions de réponses » les probabilités de répondre par l'une des trois modalités. Nous notons ici $\pi_+(T_i)$, $\pi_0(T_i)$ et $\pi_-(T_i)$ les probabilités de répondre « en hausse »(+), « stable » (0) et « en baisse » (-) avec l'évolution T_i. Par définition, la somme des fonctions $\pi_+(\cdot)$, $\pi_0(\cdot)$ et $\pi_-(\cdot)$ est égale à 1. De manière idéale, $\pi_+(\cdot)$ est une fonction croissante, nulle en $-\infty$ et égale à 1 en $+\infty$ (cf. Fig. 1). Symétriquement, $\pi_-(\cdot)$ est une fonction décroissante, égale à 1 en $-\infty$ et nulle en $+\infty$. Enfin, $\pi_0(\cdot)$ s'annule en $+\infty$ et en $-\infty$ et est maximale autour de 0.

FIG. 1 – Probabilités de réponses en fonction de l'évolution T_i (en %), fonctions de réponses idéales, réelles et dégénérées



Avec l'exemple d'une enquête de conjoncture effectuée en Suisse sur les évolutions d'investissement, Ronning (1985) a montré que ce schéma n'est pas nécessairement vérifié. Les comportements de réponses peuvent être relativement compliqués. Nous présentons schématiquement dans la Fig. 1 un exemple de fonctions de réponses estimées dans cette étude.

On suppose ici un comportement très stylisé fréquemment utilisé dans les modèles. Les entreprises répondent au regard des évolutions (prévues ou réalisées) de la variable d'intérêt en fonction de deux seuils psychologiques s^+ et s^-. Lorsque l'évolution est (ou serait) supérieure à s^+, l'entreprise répond par la modalité « en hausse », entre s^+ et s^- elle répond par la modalité « stable » et inférieure à s^-, elle répond par la modalité « en baisse ». Par simplicité, ces deux seuils sont considérés comme communs à toute la strate et constants dans le temps. Les fonctions de réponses sont alors dites « dégénérées » (cf. Fayolle, 1987). On retient alors :

$$\pi_+(T_i) = \begin{cases} 0 & \text{si} \quad T_i \leq s^+ \\ 1 & \text{si} \quad T_i > s^+ \end{cases}$$

$$\pi_0(T_i) = \begin{cases} 0 & \text{si} \quad T_i < s^- \text{ ou } T_i > s^+ \\ 1 & \text{si} \quad T_i \in [s^-, s^+] \end{cases}$$

$$\pi_+(T_i) = \begin{cases} 0 & \text{si} \quad T_i \geq s^- \\ 1 & \text{si} \quad T_i < s^- \end{cases}$$

On note alors A_+, A_0 et A_- l'espérance des évolutions de T_i selon la réponse de l'entreprise i :

$$\begin{cases} A_+ &= E[T_i|s_i = \ 1] &= E[T_i|T_i > s^+] \\ A_0 &= E[T_i|s_i = \ 0] &= E[T_i|T_i \in [s^-, s^+]\,] \\ A_- &= E[T_i|s_i = -1] &= E[T_i|T_i < s^-] \end{cases}$$

En choisissant p_i proportionnel à q_i, on dispose alors de réponses pondérées proportionnellement à la variable d'intérêt. Les parts proportionnelles à la variable d'intérêt seront notées par la suite $\hat{R}^+$, $\hat{R}^0$ et $\hat{R}^-$. Nous appellerons $\hat{S}$ le solde pondéré correspondant.

$$E[T^{\mathcal{E}}|\mathcal{I}_{-1}, \{s_i\}_i] = A_+\hat{R}^+ + A_0\hat{R}^0 + A_-\hat{R}^-$$

Détaillons davantage les valeurs de A_+, A_0 et A_- :

$$\begin{cases} A_+ &= \frac{E[T_i \mathbf{1}_{T_i > s^+}]}{P(T_i > s^+)} \\ A_0 &= \frac{E[T_i \mathbf{1}_{T_i \in [s^-, s^+]}]}{P(T_i \in [s^-, s^+])} \\ A_- &= \frac{E[T_i \mathbf{1}_{T_i < s^-}]}{P(T_i < s^-)} \end{cases}$$

$$\left\{\begin{array}{lcl} A_+ & = & \frac{\int_{s^+}^{+\infty} t f_T(t)\,dt}{\int_{s^+}^{+\infty} f_T(t)\,dt} \\ A_0 & = & \frac{\int_{s^-}^{s^+} t f_T(t)\,dt}{\int_{s^-}^{s^+} f_T(t)\,dt} \\ A_- & = & \frac{\int_{-\infty}^{s^-} t f_T(t)\,dt}{\int_{-\infty}^{s^-} f_T(t)\,dt} \end{array}\right.$$

On définit par sa densité f_0 et par sa fonction de répartition F_0 la loi centrée-réduite, transformation affine de la loi aléatoire suivie par les T_i.

$$\left\{\begin{array}{lcl} A_+ & = & m_T + \sigma_T\, a_+ \\ A_0 & = & m_T + \sigma_T\, a_0 \\ A_- & = & m_T + \sigma_T\, a_- \end{array}\right.$$

$$\text{avec} \left\{\begin{array}{lcl} a_+ & = & \frac{\int_{(s^+-m_T)/\sigma_T}^{+\infty} \tau f_0(\tau)\,d\tau}{\int_{(s^+-m_T)/\sigma_T}^{+\infty} f_0(\tau)\,d\tau} \\ a_0 & = & \frac{\int_{(s^--m_T)/\sigma_T}^{(s^+-m_T)/\sigma_T} \tau f_0(\tau)\,d\tau}{\int_{(s^--m_T)/\sigma_T}^{(s^+-m_T)/\sigma_T} f_0(\tau)\,d\tau} \\ a_- & = & \frac{\int_{-\infty}^{(s^--m_T)/\sigma_T} \tau f_0(\tau)\,d\tau}{\int_{-\infty}^{(s^--m_T)/\sigma_T} f_0(\tau)\,d\tau} \end{array}\right.$$

D'où $E[T^{\mathcal{E}}|\mathcal{I}_{-1}, \{s_i\}_i] = m_T + \sigma_T[a_+\hat{R}^+ + a_0\hat{R}^0 + a_-\hat{R}^-]$

L'introduction du solde d'opinion nécessite alors de faire deux hypothèses. On suppose d'une part que la loi des T_i est symétrique autour de sa moyenne et d'autre part que la moyenne des évolutions individuelles m_T est au centre des deux seuils psychologiques s^+ et s^-. On obtient alors $a_0 = 0$ et $a_+ = -a_-$. La formule ci-dessus devient :

$$E[T^{\mathcal{E}}|\mathcal{I}_{-1}, \{s_i\}_i] = m_T + \sigma_T\, a_+[\hat{R}^+ - \hat{R}^-]$$

Soit $\delta_T = (s^+ - m_T)/\sigma_T$ (et $\delta_T = -(s^- - m_T)/\sigma_T$). δ_T mesure l'écart des seuils

psychologiques s^+ et s^- par rapport à l'amplitude σ_T des variations individuelles. δ_T décroît avec la sensibilité des réponses s_i aux variations individuelles T_i. On peut alors définir :

$$\eta_0(\delta_T) = a_+ = -a_- = \frac{\int_{\delta_T}^{+\infty} \tau f_0(\tau) d\tau}{\int_{\delta_T}^{+\infty} f_0(\tau) d\tau}$$

Afin d'avoir un ordre de grandeur de $\eta_0(\delta_T)$, on réalise un développement limité autour de 0. Ainsi $\eta_0(\delta_T) = 2f_0(0)[1 + 2f_0(0)\delta_T] + o(\delta_T)$. Dans le cas d'une loi normale, ceci donne $\eta_0(\delta_T) = \sqrt{2/\pi} + 2/\pi \times \delta_T + o(\delta_T)$. Numériquement, ceci correspond à $\eta_0(\delta_T) = 0{,}80 + 0{,}64\delta_T + o(\delta_T)$. $\eta_0(\delta_T)$ est donc de l'ordre de 1.

Au final, la prévision *in sample* est ainsi donnée par :

$$E[T^{\mathcal{E}} | \mathcal{I}_{-1}, \{s_i\}_i] = m_T + \sigma_T \, \eta_0(\delta_T) \, \hat{S} \tag{1}$$

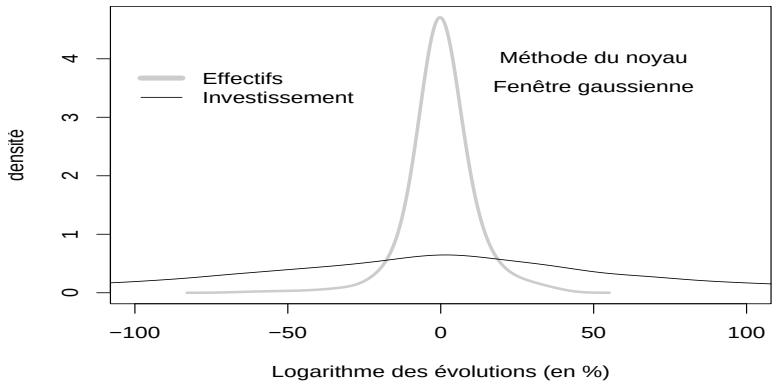
Revenons quelque peu sur les deux hypothèses nécessaires à l'introduction des soldes d'opinion. Tout d'abord, remarquons que ces hypothèses sont nécessaires à l'introduction du solde mais qu'elles ne sont que des simplifications d'écriture en ce qui concerne le choix des pondérations. Les conclusions sur ce point sont indépendantes de ces hypothèses. Ceci restera vrai tout au long de l'article.

La symétrie n'est pas naturelle. En effet, il est communément admis que les évolutions individuelles suivent des distributions proches de lois log-normales[10]. Ceci apparaît nettement en traçant les fonctions de densité des logarithmes des évolutions individuelles lorsque l'on dispose de données quantitatives (cf. Fig. 2). La distribution des T_i n'est alors pas symétrique. Toutefois, dans la mesure où cette distribution est suffisamment concentrée, elle se rapproche d'une loi normale et la symétrie peut alors être admise. L'hypothèse de symétrie peut donc être supposée vérifiée dans le cas où les variations individuelles sont faibles. Par exemple, pour les variations des effectifs salariés, la distribution des T_i est bien symétrique

[10] $log(1 + T_i)$ suit alors une loi normale.

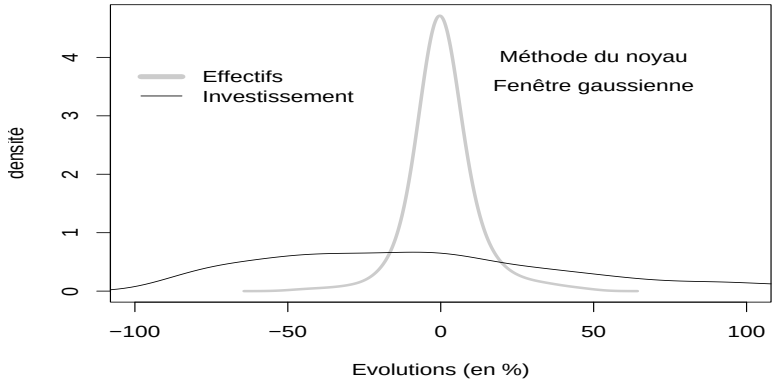
(cf. Fig. 3). Toutefois, pour une variable comme l'investissement des entreprises où les variations peuvent être très hétérogènes, la symétrie semble une hypothèse assez forte.

FIG. 2 – Densité des logarithmes des évolutions annuelles individuelles des effectifs salariés et de l'investissement



Source : Répertoire FUTE, Insee. Calculs des auteurs.
Note : FUTE est un répertoire qui utilise à la fois les données issues des Enquêtes Annuelles d'Entreprises (EAE) et des données fiscales.

FIG. 3 – Densité des taux de croissance annuels individuels des effectifs salariés et de l'investissement



Source : Répertoire FUTE, Insee. Calculs des auteurs.

L'hypothèse de symétrie n'étant pas nécessairement vérifiée en pratique, un modèle équivalent est développé en annexe A à l'aide de lois log-normales. Cette approche avait été retenue dans Fansten (1976). En annexe A, le modèle est écrit sous forme multiplicative. Les conclusions sont très proches du modèle additif développé dans le corps du texte mais l'expression littérale est un peu plus compliquée.

L'hypothèse $m_T = 1/2\,(s^+ + s^-)$ est quant à elle beaucoup plus restrictive. Elle n'est pas nécessairement contre-intuitive, mais rien n'oblige à ce qu'elle soit valide. Notons qu'elle était également présente de manière implicite dans Fansten (1976)[11]. D'expérience, on sait qu'un solde d'opinion peut rester de signe constant, ceci quelle que soit la situation conjoncturelle. En particulier, pour les soldes qui restent tout le temps négatifs, on peut penser que m_T est significativement inférieure au centre des deux seuils psychologiques s^+ et s^-.

Le solde d'opinion apparaît alors comme un résumé partiel de l'information contenue dans les proportions de réponses. Cette hypothèse a été étudiée par Hild (2002), qui montre qu'il est en effet possible de mieux prévoir les évolutions réelles en utilisant à la fois les soldes d'opinion et les proportions de réponses « stable ».

3.2 Pondérations optimales pour l'estimation *out of sample*

Il n'y pas de lien comptable entre les évolutions individuelles T_i de l'échantillon $\mathcal{E}$ et l'évolution de l'agrégat hors échantillon $\mathcal{P} \setminus \mathcal{E}$. Toutefois l'échantillon est construit de manière à être représentatif de la population toute entière. Dans ce cas, on peut admettre qu'il y a homogénéité entre les comportements des entreprises de $\mathcal{E}$ et de $\mathcal{P} \setminus \mathcal{E}$. La stratification renforce cette homogénéité. Les strates sont construites de manière à assurer au mieux la similitude des comportements. Au final, il est possible de considérer que, par strate, les comportements économiques des entreprises sont en grande partie déterminées par une situation conjoncturelle

[11] Dans Fansten (1976), les développements limités permettant l'introduction du solde d'opinion nécessitent d'avoir (dans les notations de l'article) $\sqrt{s^+/s^-} \approx i_m$. Dans les notations de notre modèle additif, ceci s'écrit $m_T \approx 1/2\,(s^+ + s^-)$.

commune.

Supposons alors qu'une entreprise i, appartenant ou non à l'échantillon, connaisse un taux de croissance T_i donné par $Z + \epsilon_i$. Z est une variable conjoncturelle sous-jacente qui est commune à toute la population (de la strate). Z est distribuée selon une loi aléatoire, identique au cours du temps, de moyenne m_Z et de variance σ_Z^2. Sa fonction de répartition est notée F_Z et sa densité f_Z. Par simplification, les Z sont considérés dans cette étude comme non-autocorrélés. Cette hypothèse n'est pas gênante puisque nous recherchons une relation instantanée entre les réponses des entreprises et les évolutions réelles de l'agrégat. Les ϵ_i sont considérés comme des variables aléatoires indépendantes et identiquement distribuées de moyenne nulle et de variance σ_ϵ^2. Les ϵ_i sont indépendants de Z. De la même manière, la recherche d'une relation instantanée nous invite à considérer les ϵ_i comme indépendants dans le temps. Leur fonction de répartition est F_ϵ et leur densité f_ϵ. Par construction, on a $m_Z = m_T$ et $\sigma_T^2 = \sigma_Z^2 + \sigma_\epsilon^2$.

Par linéarité et par indépendance des mouvements idiosyncrasiques ϵ_i, on peut écrire :

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}} \,|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = E\left[T_j \,|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] \text{ avec } j \in \mathcal{P}\setminus\mathcal{E}$$

Comme les ϵ_i sont d'espérance nulle, on en déduit la formule (2) ci-après. Elle exprime l'idée naturelle que la prévision des évolutions *out of sample* se résume à prévoir celle de la variable sous-jacente Z.

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}} \,|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = E\left[Z \,|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] \tag{2}$$

La formule de Bayes donne la densité de la loi de Z connaissant les réalisations $\{s_i\}$ des réponses des entreprises de l'échantillon :

$$f\left(z|\{s_i\}\right) = \frac{P\left(\{S_i\} = \{s_i\}|Z = z\right) \cdot f_Z(z)}{P\left(\{S_i\} = \{s_i\}\right)}$$

On a naturellement :

$$P\left(\{S_i\} = \{s_i\}\right) = \int_{-\infty}^{+\infty} P\left(\{S_i\} = \{s_i\} | Z = z\right) \cdot f_Z(z)\, dz$$

$$\text{D'où :}\ \ E\left[Z | \mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = \frac{\int_{-\infty}^{+\infty} z \cdot P\left(\{S_i\} = \{s_i\} | Z = z\right) \cdot f_Z(z)\, dz}{\int_{-\infty}^{+\infty} P\left(\{S_i\} = \{s_i\} | Z = z\right) \cdot f_Z(z)\, dz} \quad (3)$$

$$\begin{aligned} \text{et :}\ \ P\left(\{S_i\}_{i\in\mathcal{E}} = \{s_i\}_{i\in\mathcal{E}} | Z = z\right) &= \left(1 - F_\epsilon(s^+ - z)\right)^{N^+} \\ &\times \left(F_\epsilon(s^+ - z) - F_\epsilon(s^- - z)\right)^{N^0} \\ &\times F_\epsilon(s^- - z)^{N^-} \end{aligned} \quad (4)$$

Dans la formule (4), N^+, N^0 et N^- désignent le nombre de réponses « en hausse », « stable » et « en baisse ». Les formules (3) et (4) donnent le résultat central : les parts non pondérées de réponse dans chacune des trois modalités résument parfaitement l'information fournie par les réponses des entreprises pour prévoir la variable conjoncturelle sous-jacente Z. Elles contiennent donc parfaitement toute l'information donnée par les entreprises de l'échantillon afin de prévoir la variation agrégée $T^{\mathcal{P}\setminus\mathcal{E}}$ de la variable d'intérêt dans la population en dehors de l'échantillon.

Rappelons toutefois que ce résultat n'est vrai que sous couvert que les strates soient construites de manière à ce que les comportements soient suffisamment homogènes au sein de chacune d'entre elles. En particulier, lorsqu'il n'y a pas de stratification par taille d'entreprise, il est probable que les comportements conjoncturels des entreprises soient fonction de leur taille. Nous sortons alors du cadre du modèle. Dans ce cas, pour les prévisions *out of sample*, des pondérations croissantes avec la taille des entreprises sont susceptibles d'être meilleures que des pondérations constantes.

L'espérance de Z n'est pas calculable directement. Il est alors nécessaire de faire quelques simplifications afin de justifier l'utilisation du solde d'opinion non pon-

déré.

On linéarise (4) en faisant un développement limité en $z - m_Z$. Ceci est légitime pour des distributions très concentrées de Z. Les taux de croissance trimestriels ou mensuels ont rarement des écarts-types supérieurs à 10% et il est donc possible de considérer $Z - m_Z$ comme négligeable devant 1.

Pour simplifier l'écriture, on note :

$$\begin{cases} F_+ = F_\epsilon(s^+ - m_Z) \\ F_- = F_\epsilon(s^- - m_Z) \\ f_+ = f_\epsilon(s^+ - m_Z) = F'_\epsilon(s^+ - m_Z) \\ f_- = f_\epsilon(s^- - m_Z) = F'_\epsilon(s^- - m_Z) \end{cases}$$

$$\begin{aligned} P\left(\{S_i\}_{i\in\mathcal{E}} = \{s_i\}_{i\in\mathcal{E}} | Z = z\right) &= (1-F_+)^{N^+} \times (F_+ - F_-)^{N^0} \times F_-^{N^-} \\ &\times \left[1 + N\tilde{\Gamma}(z - m_Z) + (N\tilde{\Phi} + N^2\tilde{\Omega})(z - m_Z)^2\right] \\ &+ o((z - m_Z)^2) \end{aligned}$$

avec :

$$\begin{cases} \tilde{\Gamma} &= \frac{f_+}{1-F_+}\tilde{R}^+ - \frac{f_+ - f_-}{F_+ - F_-}\tilde{R}^0 - \frac{f_-}{F_-}\tilde{R}^- \\ \tilde{\Phi} &= \frac{1}{2}\tilde{R}^-\left(\frac{f'_-}{F_-} - \frac{f_-^2}{F_-^2}\right) + \frac{1}{2}\tilde{R}^0\left(\frac{f'_+ - f'_-}{F_+ - F_-} - \frac{(f_+ - f_-)^2}{(F_+ - F_-)^2}\right) + \frac{1}{2}\tilde{R}^+\left(-\frac{f'_+}{1-F_+} - \frac{f_+^2}{(1-F_+)^2}\right) \\ \tilde{\Omega} &= \tilde{R}^-\tilde{R}^0\frac{f_-}{F_-}\frac{f_+ - f_-}{F_+ - F_-} - \tilde{R}^0\tilde{R}^+\frac{f_+ - f_-}{F_+ - F_-}\frac{f_+}{1-F_+} - \tilde{R}^+\tilde{R}^-\frac{f_+}{1-F_+}\frac{f_-}{F_-} \\ &\quad + \frac{1}{2}\left(\tilde{R}^-\right)^2\frac{f_-^2}{F_-^2} + \frac{1}{2}\left(\tilde{R}^0\right)^2\frac{(f_+ - f_-)^2}{(F_+ - F_-)^2} + \frac{1}{2}\left(\tilde{R}^+\right)^2\frac{f_+^2}{(1-F_+)^2} \end{cases}$$

$\tilde{R}^+$, $\tilde{R}^0$ et $\tilde{R}^-$ désignent les proportions non pondérées de chaque type de réponse. On a bien sûr les relations $\tilde{R}^+ = N^+/N$, $\tilde{R}^0 = N^0/N$ et $\tilde{R}^- = N^-/N$. Le solde correspondant, non pondéré, est noté $\tilde{S}$. En utilisant la formule (3), on obtient

alors :

$$E\left[Z|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]=m_Z+\frac{\sigma_Z^2N\tilde{\Gamma}+o(\sigma_Z^2)}{1+(N\tilde{\Phi}+N^2\tilde{\Omega})\sigma_Z^2+o(\sigma_Z^2)} \tag{5}$$

Sous l'hypothèse que $(N\tilde{\Phi}+N^2\tilde{\Omega})\sigma_Z^2$ soit négligeable devant 1, on peut faire l'approximation au premier ordre :

$$E\left[Z|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]=m_Z+\sigma_Z^2N\tilde{\Gamma}+o(\sigma_Z^2) \tag{6}$$

soit encore

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]=m_T+\sigma_Z^2N\tilde{\Gamma}+o(\sigma_Z^2) \tag{7}$$

Cette hypothèse est vérifiée dans le cas qui nous intéresse ici où les variations idiosyncrasiques (de variance σ_ϵ^2) sont suffisamment importantes devant les variations communes (de variance σ_Z^2), ceci d'un facteur très supérieur à un facteur N[12].

En ayant choisi des pondérations individuelles constantes, le taux de croissance sur la population hors échantillon est alors estimé par une fonction des proportions de réponse « en hausse », « stable » et « en baisse ». En première approximation, cette fonction est une fonction affine de ces proportions.

Si on se place dans le cas particulier où la distribution des ϵ_i est symétrique et où les variations sont centrées autour du centre des deux seuils psychologiques ($m_T=1/2\left(s^+ + s^-\right)$), on est alors dans le cas particulier où $\tilde{\Gamma}\equiv\frac{f_-}{F_-}\tilde{S}=\frac{f_+}{1-F_+}\tilde{S}$. Dans ce cas particulier, en première approximation, le taux de croissance sur la population hors échantillon est estimé par une fonction affine du solde non pondéré $\tilde{S}$.

[12]Nous verrons par la suite que c'est en particulier le cas pour les questions relatives à l'emploi dans l'enquête Activité dans l'industrie de l'Insee.

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right] = m_T + \sigma_Z^2 N \frac{f_\epsilon(\frac{s^+-s^-}{2})}{1-F_\epsilon(\frac{s^+-s^-}{2})}\tilde{S} + o(\sigma_Z^2) \qquad (8)$$

On défini f_1 et F_1 les fonctions de densité et de répartition des $\epsilon_i/\sigma_\epsilon$. La formule (8) devient alors :

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right] = m_T + \sigma_Z N \frac{\sigma_Z}{\sigma_\epsilon}\eta_1(\delta_\epsilon)\tilde{S} + o(\sigma_Z^2) \qquad (9)$$

$$\text{avec } \eta_1(\delta_\epsilon) = \frac{f_1(\delta_\epsilon)}{1-F_1(\delta_\epsilon)} \qquad (10)$$

$$\text{et } \delta_\epsilon = \frac{s^+-s^-}{2\sigma_\epsilon} = \frac{s^+-m_T}{\sigma_\epsilon} = \frac{\sigma_T}{\sigma_\epsilon}\delta_T$$

Le facteur σ_Z est un facteur d'échelle. Le facteur σ_Z/σ_ϵ, noté $1/\rho$ par la suite, indique que la prévision est d'autant plus sensible au solde que les évolutions individuelles sont homogènes. Autrement dit, le solde estime bien les variations de la composante commune si celles-ci sont grandes par rapport aux mouvements idiosyncrasiques. Ensuite, la sensibilité croît avec la taille N de l'échantillon. En effet, plus l'échantillon est grand, plus la composante commune est bien estimée. Enfin, le terme $\eta_1(\delta_\epsilon)$ décroît lorsque l'écart augmente entre les deux seuils s^- et s^+ (écart mesuré relativement à σ_ϵ par δ_ϵ).

On cherche à fixer un ordre de grandeur aux différents paramètres de la relation. Comme pour $\eta_0(\delta_T)$ dans la partie 3.1, $\eta_1(\delta_\epsilon)$ est de l'ordre de 1[13]. Dans l'enquête Activité dans l'industrie de l'Insee[14], pour les questions « produits », il y a en moyenne 53 unités (entreprises $\times$ produits) répondantes par strate. Pour les ques-

[13]La linéarisation autour de 0 donne $\eta_1(\delta_\epsilon) = \frac{f_1(0)}{1-F_1(0)}[1 + (\frac{f_1'(0)}{f_1(0)} + \frac{f_1(0)}{1-F_1(0)})\delta_\epsilon] + o(\delta_\epsilon)$. Dans le cas de la loi normale, ceci s'écrit $\eta_1(\delta_\epsilon) = \sqrt{\frac{2}{\pi}} + \frac{2}{\pi}\delta_\epsilon + o(\delta_\epsilon)$, soit encore $\eta_1(\delta_\epsilon) = 0,80 + 0,64\,\delta_\epsilon + o(\delta_\epsilon)$.

[14]Dans cette enquête (cf. 5.1), deux types de questions sont posées : d'une part des questions relatives aux produits de l'entreprises (évolution de la production, des carnets de commandes, ... pour chacun des produits renseignés) et d'autre part, des questions qui portent sur l'entreprise dans son ensemble (emploi, goulots de production, ...).

tions « entreprise », il y a en moyenne 13 entreprises répondants par strate. Pour fixer les esprits, nous posons $N = 50$. D'autre part, nous choisissons $\sigma_Z = 0{,}5\%$ et $\sigma_\epsilon = 5\%$. Dans ce cas, le coefficient multiplicatif[15] est d'environ 2,5%. Une différence de 10 points du solde indique un écart de $0,25$ point du taux de croissance. Ceci semble un ordre de grandeur très raisonnable.

3.3 Pondérations optimales pour l'agrégat global, modèle simplifié

Nous avons montré (partie 3.1) que, sous les hypothèses retenues, la prévision *in sample* est proportionnelle au solde des réponses pondérées proportionnellement à la taille des entreprises (partie 3.1). Au contraire, la prévision *out of sample* est proportionnelle au solde non pondéré (partie 3.2). Il est alors possible d'obtenir un solde prévoyant les évolutions globales par une combinaison linéaire du solde pondéré et du solde non pondéré.

$$E\left[T|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right] = QE\left[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right] + (1-Q)E\left[T^{\mathcal{P}\setminus\mathcal{E}}|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]$$

En utilisant $\sigma_T^2 = \sigma_Z^2 + \sigma_\epsilon^2$ et $\rho = \sigma_\epsilon/\sigma_Z$, les formules (1) et (9) s'écrivent :

$$\left\{\begin{array}{lcl} E[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_j\}_{j\in\mathcal{E}}] & = & m_T + \sigma_Z\left\{\rho\sqrt{1+1/\rho^2}\eta_0(\delta_T)\right\}\hat{S} + o\left(\sigma_Z^2\right) \\ E\left[T^{\mathcal{P}\setminus\mathcal{E}}|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right] & = & m_T + \sigma_Z\left\{\frac{1}{\rho}\eta_1(\delta_\epsilon)N\right\}\tilde{S} + o\left(\sigma_Z^2\right) \end{array}\right.$$

L'agrégation donne alors :

$$\begin{array}{rcl} E[T|\mathcal{I}_{-1},\{s_j\}_{j\in\mathcal{E}}] & = & m_T \\ & & +\sigma_Z\left\{Q\,\rho\sqrt{1+1/\rho^2}\eta_0(\delta_T)\right\}\hat{S} \\ & & +\sigma_Z\left\{(1-Q)\,\frac{1}{\rho}\eta_1(\delta_\epsilon)N\right\}\tilde{S} \\ & & +o\left(\sigma_Z^2\right) \end{array}$$

Le poids relatif à affecter au solde pondéré par rapport au solde non pondéré est

[15] i.e. $\sigma_Z N\frac{\sigma_Z}{\sigma_\epsilon}\eta_1(\delta_\epsilon)$

alors donné par la formule :

$$\frac{\lambda}{1-\lambda} = \frac{Q}{1-Q} \frac{\frac{\eta_0(\delta_T)}{\eta_1(\delta_\epsilon)} \sqrt{1+1/\rho^2}\, \rho^2}{N} \tag{11}$$

Dit autrement, avec la formule (11), le solde optimal serait obtenu en utilisant les poids individuels p_i donnés par :

$$p_i = \lambda \frac{q_i}{Q} + (1-\lambda)\frac{1}{N}$$

Le poids à affecter au solde pondéré est donc croissant avec le taux de couverture Q. Lorsque le taux de couverture est de 100%, les pondérations optimales sont les pondérations proportionnelles à la taille des entreprises. Les enquêtes de conjoncture de l'Insee présentent toutes une strate exhaustive correspondant aux plus grandes entreprises. En l'absence des non réponses, il faudrait donc utiliser des pondérations proportionnelles pour la strate exhaustive.

Toutefois, nous avons fait une simplification importante dans la partie 3.1, où nous avons considéré que pour prévoir la variation individuelle X_i d'une entreprise i répondant à l'enquête, les réponses des autres entreprises de l'échantillon n'apportent pas d'information. Ceci revient à négliger la composante conjoncturelle Z, commune aux différentes entreprises de l'échantillon $\mathcal{E}$. En levant cette hypothèse (partie 4), quel que soit le taux de couverture Q, les pondérations proportionnelles donnent trop d'importance à la taille des entreprises. En particulier, ceci reste vrai même dans le cas extrême d'un taux de couverture de 100% correspondant à la strate exhaustive.

4 Modèle plus complet : les prévisions *in sample* dépendent des évolutions conjoncturelles sous-jacentes

4.1 Les pondérations optimales pour les prévisions *in sample*, modèle plus complet

Dans la partie 3.1, on a considéré que les réponses $\{s_j\}_{j\in\mathcal{E}\setminus\{i\}}$ des autres entreprises j de l'échantillon $\mathcal{E}$ n'influaient pas sur la prévision des évolutions T_i relatives à l'entreprise i de l'échantillon. En réalité, cette hypothèse est assez restrictive. Supposons, pour fixer les esprits, que l'entreprise i répond « en hausse ». Si la situation conjoncturelle est favorable (Z important), l'évolution T_i est susceptible d'être très forte. Au contraire, si la situation conjoncturelle est dégradée, il est probable que la hausse de T_i reste limitée.

Par la suite, nous utilisons de ce fait pour les prévisions *in sample* le modèle développé dans la partie 3.2 (dans le cadre des prévisions *out of sample*). Dans ce cas, T_i est alors égale à $Z + \epsilon_i$.

$$E[T_i|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] = E[Z|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] + E[\epsilon_i|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] \qquad (12)$$

Le premier terme est donné par la formule (6) de la partie 3.2 (prévisions *out of sample*). Le deuxième terme s'écrit :

$$E[\epsilon_i|\{s_j\}_{j\in\mathcal{E}}] = E_Z\left[\, E_\epsilon[\epsilon_i|Z=z, \{s_j\}_{j\in\mathcal{E}}] \,|\{s_j\}_{j\in\mathcal{E}}\right]$$

ϵ_i est idiosyncrasique. À z donné, les $\{s_j\}_{j\in\mathcal{E}\setminus\{i\}}$ n'apportent donc aucune information pour la prévision de ϵ_i. L'expression ci-dessus peut donc se réécrire :

$$E[\epsilon_i|\{s_j\}_{j\in\mathcal{E}}] = E_Z\left[\, E_\epsilon[\epsilon_i|Z=z, s_i] \,|\{s_j\}_{j\in\mathcal{E}}\right]$$

Les calculs intermédiaires sont donnés dans l'annexe B. Pour s appartenant à $\{+, 0, -\}$ (correspondant aux réponses « hausse », « stable » et « baisse »), on définit C_s, D_s et G_s. Leurs expressions sont données dans cette même annexe. On

obtient alors :

$$\left\{\begin{array}{rcl} E_\epsilon[\epsilon_i|Z=z,s_i=-1] & = & C_-\left\{1+D_-(z-m_Z)+G_-(z-m_Z)^2\right\}+o((z-m_Z)^2) \\ E_\epsilon[\epsilon_i|Z=z,s_i=\ \ 0] & = & C_0\left\{1+D_0(z-m_Z)+G_0(z-m_Z)^2\right\}+o((z-m_Z)^2) \\ E_\epsilon[\epsilon_i|Z=z,s_i=\ \ 1] & = & C_+\left\{1+D_+(z-m_Z)+G_+(z-m_Z)^2\right\}+o((z-m_Z)^2) \end{array}\right.$$

Ces trois expressions permettent de calculer l'espérance de T_i en utilisant l'ensemble des réponses $\{s_j\}_{j\in\mathcal{E}}$ des entreprises de l'échantillon (cf. annexe B) :

$$E[T_i|\{s_j\}_{j\in\mathcal{E}},s_i=s]=m_T+C_s(1+G_s\sigma_Z^2)+N\tilde{\Gamma}\sigma_Z^2+C_sD_sN\tilde{\Gamma}\sigma_Z^2+o\left(\sigma_Z^2\right)$$

On utilise $T^\epsilon=\Sigma_{i\in\mathcal{E}}\frac{q_i}{Q}T_i$ avec $\Sigma_{i\in\mathcal{E}}\frac{q_i}{Q}=1$. L'agrégation donne alors :

$$\begin{array}{rl} E[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_j\}_{j\in\mathcal{E}}]= & m_T \\ & +\sum_{s\in\{+,0,-\}}\hat{R}^sC_s(1+G_s\sigma_Z^2) \\ & +N\tilde{\Gamma}\sigma_Z^2 \\ & +\left(\sum_{s\in\{+,0,-\}}\hat{R}^sC_sD_s\right)N\tilde{\Gamma}\sigma_Z^2 \\ & +o\left(\sigma_Z^2\right) \end{array} \qquad (13)$$

Pour revenir aux notations de la partie 3.1, on calcule A_+, A_0 et A_- avec la modélisation de cette partie ($T_i=Z+\epsilon_i$).

$$\forall s\in\{+,0,-\},\ A_s=E\left[Z+\epsilon_i|s_i=s\right]$$

Il est alors possible d'écrire (cf. annexe B) :

$$\begin{array}{rl} E[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_j\}_{j\in\mathcal{E}}]= & m_T \\ & +\sigma_T\sum_{s\in\{+,0,-\}}a_s\hat{R}^s \\ & +N\tilde{\Gamma}\sigma_Z^2 \\ & +\left(\sum_{s\in\{+,0,-\}}\hat{R}^sC_sD_s\right)N\tilde{\Gamma}\sigma_Z^2 \\ & +o\left(\sigma_Z^2\right) \end{array}$$

A la moyenne m_T, on ajoute trois termes. Le premier correspond à la formule obtenue pour les prévisions *in sample* dans la partie 3.1. Il ne fait intervenir que les proportions pondérées proportionnellement. Le second terme correspond à la formule retenue pour les estimations *out of sample* (partie 3.2). Il est fonction des proportions non pondérées. Enfin, le troisième terme apparaît comme un terme correctif qui croise les proportions pondérées et non-pondérées.

Cette expression est plus complexe que celle obtenue dans la partie 3. En particulier, elle ne permet pas d'écrire la prévision $E[T]$ sur la population toute entière comme une combinaison linéaire des soldes pondérés et des soldes non pondérés. Toutefois, elle indique clairement que les soldes non pondérés interviennent également pour estimer les évolutions *in sample*. En levant l'hypothèse que les réponses des autres entreprises (entreprise de $\mathcal{E} \setminus \{i\}$) n'apportent pas d'information pour estimer l'évolution relative à l'entreprise i de l'échantillon $\mathcal{E}$, les soldes non pondérés interviennent alors dans l'estimation des variations conjoncturelles communes.

Comme dans la partie précédente (partie 3), on fait l'hypothèse de la symétrie de s^+ et de s^- autour de m_T et on considère comme symétrique la distribution des ϵ_i. Sous ces deux hypothèses, on obtient $I_+ = -I_- > 0$, $I_0 = 0$, $C_+ = -C_- = C > 0$, $C_0 = 0$, $D_- = -D_+ = D > 0$, $D_0 = 0$ et $G_- = G_+ = G$. Le signe de G est indéterminé en général. De manière empirique, il apparaît être positif.

A partir de la loi des ϵ normalisée - de fonctions de répartition et de densité F_1 et f_1 - on définit les fonctions :

$$\begin{cases} I_1(\delta) = \int_\delta^{+\infty} e f_1(e) de \\ c_1(\delta) = \frac{I_1(\delta)}{1-F_1(\delta)} \\ d_1(\delta) = \frac{f_1(\delta)}{1-F_1(\delta)} - \frac{\delta f_1(\delta)}{I_1(\delta)} \\ g_1(\delta) = -\frac{1}{2}\frac{\delta f_1(\delta)' + f_1(\delta)}{I_1(\delta)} - \frac{\delta f_1(\delta)^2}{I_1(\delta)(1-F_1(\delta))} + \frac{1}{2}\frac{f_1(\delta)'}{1-F_1(\delta)} + \frac{f_1(\delta)^2}{(1-F_1(\delta))^2} \end{cases}$$

Avec ces notations, on obtient ainsi :

$$\left\{ \begin{array}{l} C = \sigma_\epsilon c_1(\delta_\epsilon) \\ D = \frac{1}{\sigma_\epsilon} d_1(\delta_\epsilon) \\ G = \frac{1}{\sigma_\epsilon^2} g_1(\delta_\epsilon) \end{array} \right.$$

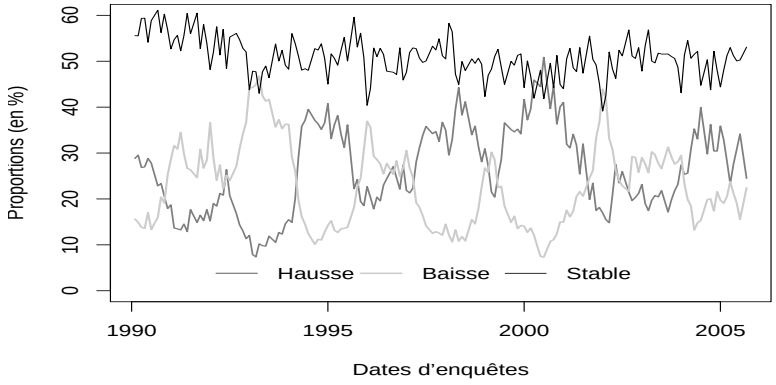
L'équation (13) devient alors :

$$\begin{array}{rl} E[T^{\mathcal{E}}|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] = & m_T \\ & +\sigma_Z \left\{ \rho c_1(\delta_\epsilon) \left(1 + g_1(\delta_\epsilon) \frac{1}{\rho^2} \right) \right\} \hat{S} \\ & +\sigma_Z \left\{ \frac{1}{\rho} \eta_1(\delta_\epsilon) N \right\} \tilde{S} \\ & -\sigma_Z \left\{ \frac{1}{\rho} \eta_1(\delta_\epsilon) c_1(\delta_\epsilon) d_1(\delta_\epsilon) N \right\} \left(\hat{R}^+ + \hat{R}^- \right) \tilde{S} \\ & +o\left(\sigma_Z^2\right) \end{array} \tag{14}$$

Or, la proportion pondérée des réponses à la modalité « stable » ne varie peu au cours du temps. Par exemple, dans l'enquête sur l'activité dans l'industrie, pour la question relative aux évolutions passées de la production, la proportion pondérée de réponses à la modalité « stable » varie avec un écart-type de 4,2% alors que les proportions pondérées de réponses selon les modalités « hausse » et « baisse » varient selon des écart-types de 9,3% et de 8,8% (cf. 4). La proportion des réponses différentes à la modalité « stable » peut donc être considérée comme constante. On note alors $\hat{M}(\delta_T) \equiv \hat{R}^+ + \hat{R}^-$. Avec $\hat{M}(\delta_T) = E[\hat{R}^+ + \hat{R}^-]$, on obtient facilement $\hat{M}(\delta_T) = 2(1 - F_0(\delta_T))$. Pour donner un ordre de grandeur, $\hat{M}(\delta_T)$ est proche de 50%.[16]

[16]Ceci correspond à $\delta_T = 67\%$. La sensibilité des réponses des entreprises apparaît très faible puisqu'il faut une hausse individuelle supérieure de 67% à l'évolution moyenne pour qu'une entreprise réponde « hausse ».

FIG. 4 – Proportions de réponses par modalités, tendance passée de la production, enquête sur l'activité dans l'industrie



Source : Enquête sur l'activité dans l'industrie, Insee.

Au final, nous retenons pour les prévisions d'évolution de l'agrégat *in sample* :

$$\begin{aligned} E[T^{\mathcal{E}}|\mathcal{I}_{-1},\{s_j\}_{j\in\mathcal{E}}] = \quad & m_T \\ & +\sigma_Z\left\{\rho c_1(\delta_\epsilon)\left(1+g_1(\delta_\epsilon)\tfrac{1}{\rho^2}\right)\right\}\hat{S} \\ & +\sigma_Z\left\{\tfrac{1}{\rho}\eta_1(\delta_\epsilon)N\right\}\tilde{S} \\ & -\sigma_Z\left\{\tfrac{1}{\rho}\eta_1(\delta_\epsilon)c_1(\delta_\epsilon)d_1(\delta_\epsilon)N\hat{M}(\delta_T)\right\}\tilde{S} \\ & +o\left(\sigma_Z^2\right) \end{aligned} \tag{15}$$

4.2 Pondérations optimales pour l'agrégat global, modèle plus complet

Contrairement aux résultats de la partie 3.1, il apparait que les soldes non pondérés apparaissent également dans les prévisions *in sample*. Ceci modifie donc sensiblement les résultats de la partie 3.3. L'agrégation des deux prévisions *in sample* et *out of sample* donne désormais :

$$
\begin{aligned}
E[T|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] \;=\; & m_T \\
& +\sigma_Z \left\{ Q\,\rho c_1(\delta_\epsilon) \left(1 + g_1(\delta_\epsilon)\tfrac{1}{\rho^2}\right) \right\} \hat{S} \\
& +\sigma_Z \left\{ \tfrac{1}{\rho}\eta_1(\delta_\epsilon) N \right\} \tilde{S} \\
& -\sigma_Z \left\{ Q\,\tfrac{1}{\rho}\eta_1(\delta_\epsilon) c_1(\delta_\epsilon) d_1(\delta_\epsilon) N \hat{M}(\delta_T) \right\} \tilde{S} \\
& +o\left(\sigma_Z^2\right)
\end{aligned}
$$

Le poids relatif à affecter aux pondérations proportionnelles à la taille des entreprises par rapport aux pondérations constantes est donc donné par la formule :

$$
\frac{\lambda}{1-\lambda} = \frac{Q\frac{c_1(\delta_\epsilon)}{\eta_1(\delta_\epsilon)} \left(1 + g_1(\delta_\epsilon)\frac{1}{\rho^2}\right) \rho^2}{\left(1 - Q c_1(\delta_\epsilon) d_1(\delta_\epsilon) \hat{M}(\delta_T)\right)\; N} \tag{16}
$$

Le poids relatif λ est strictement inférieur à 1 même dans le cas d'un taux de couverture Q de 100% : des pondérations proportionnelles donnent toujours trop d'importance aux réponses des grandes entreprises de la state.[17]

La formule (16) peut être simplifiée. Tout d'abord, il est naturel de considérer que les évolutions idiosyncrasiques sont significativement plus importantes que les évolutions conjoncturelles. Ceci s'écrit $\sigma_Z^2 \ll \sigma_\epsilon^2$. D'une part, on obtient ainsi $\sigma_\epsilon \approx \sigma_T$, soit encore $\delta_\epsilon \approx \delta_T$. D'autre part, ρ^2 est alors grand devant 1. De plus, $g_1(\delta_\epsilon)$ est inférieur à 1.[18] Il est donc possible de négliger le terme $g_1(\delta_\epsilon)/\rho^2$.

On définit $h_1(\delta) = \frac{c_1(\delta)}{\eta_1(\delta)}$ et $k_1(\delta) = c_1(\delta) d_1(\delta) \hat{M}(\delta)$. D'une part, concernant le premier élément, $h_1(0)$ vaut 1 et $h_1(\delta)$ - pour δ positif et fixé - croît avec l'épaisseur des queues de distribution de la loi symétrique envisagée. Dans le cas particulier d'une loi normale, h_1 est tout le temps égale à 1. En première approximation, nous pouvons considérer que $h_1(\delta)$ est ainsi tout le temps égal à 1.

[17] Empiriquement, $c_1(\delta_\epsilon) d_1(\delta_\epsilon) \hat{M}(\delta_T)$ est strictement inférieur à 1, cf. ci-dessous.

[18] Dans le cas d'une loi normale, $g_1(0) = 13\%$ et $g_1(\delta)$ est décroissant sur $[0, +\infty[$. $g_1(\delta)$ vaut 10% pour $\delta = 20\%$, 4% pour $\delta = 1$ et $g_1(\delta)$ tend vers 0 lorsque δ tend vers l'infini.

D'autre part, concernant le deuxième élément, k_1 est décroissante sur $[0, +\infty[$. Pour donner un ordre de grandeur, dans le cas d'une loi normale, k_1 vaut 64% en 0, 57% en 20%, 50% en 38% et tend vers 0 à l'infini. Puisque Q et $k_1(\delta_\epsilon)$ sont inférieurs à 1, le dénominateur de la formule (16) est strictement positif : le poids à donner au solde non-pondéré n'est jamais nul.

Au regard de ces différents éléments, la formule (16) se réécrit :

$$\frac{\lambda}{1-\lambda} = \frac{Q\rho^2}{(1 - Qk_1(\delta_\epsilon))\ N} \tag{17}$$

N est parfaitement connu. Q peut être calculé à partir des pondérations primaires de l'échantillon et des coefficients de redressement. ρ est cependant plus difficile à connaître précisément. Dans la partie 3.2, nous avions retenu comme valeur $\rho = 10$. Pour fixer les esprits, en gardant cette valeur pour ρ, avec 100 observations dans la strate et $\delta_\epsilon = 38\%$, les poids relatifs optimaux seraient $\lambda = 40\%$ pour les soldes pondérés et $1 - \lambda = 60\%$ pour les soldes non pondérés.

– De manière naturelle, la part relative du solde pondéré doit augmenter avec le taux de couverture pondéré Q de l'enquête. Par exemple, si le taux de couverture est très proche de 0, le poids affecté au solde pondéré doit être nul. On est alors dans le cas de la partie 3.2 des prévisions *out of sample*. A l'opposé, lorsque le taux de couverture est proche de 1, le poids affecté aux soldes non pondérés est faible. Toutefois, ce poids ne peut être nul. C'est le résultat de la partie 4.1 : les soldes non pondérés interviennent dans les prévisions *in sample*.

– Les procédures de tirages aléatoires[19] dans les plans de sondages des enquêtes de conjoncture de l'Insee sont, au sein d'une strate, à probabilité constante, indépendante de la taille de l'entreprise[20]. Aussi, le rapport Q/N peut être interprété

[19] On rappelle que seules les entreprises n'appartenant pas à la strate exhaustive sont sélectionnées de façon aléatoire.

[20] En revanche, le nombre d'entreprises à sélectionner dans chaque strate est obtenu selon une allocation de Neyman, proportionnellement au poids de la strate dans l'ensemble de l'économie.

comme une mesure de la concentration de la population étudiée. Le poids accordé aux soldes pondérés est alors croissant avec la concentration.

– ρ est une mesure de l'hétérogénéité des évolutions individuelles[21]. Le poids accordé aux soldes pondérés augmente fortement (de manière quadratique) avec l'importance des évolutions idiosyncrasiques. Ceci est naturel puisque, dans le cas où les évolutions conjoncturelles n'expliquent qu'une faible partie des évolutions individuelles, la prévision *out of sample* est peu précise. Ceci signifie par exemple que le poids accordé aux soldes pondérés doit être plus élevé dans le cas des évolutions d'investissement que dans le cas des évolutions de la production.

4.3 Que faire en pratique ?

Dans la réalité, il n'est pas possible de connaître les différents paramètres. En particulier, ρ - indicateur d'hétérogénéité - est difficilement mesurable. Il apparaît toutefois que pondérer les soldes proportionnellement ou ne pas les pondérer sont des comportements polaires.

Il est bien entendu possible de choisir entre ces deux alternatives. Les éléments développés ci-dessus permettent d'éclairer de manière théorique ce choix dans la mesure où des stratégies intermédiaires de pondérations sont envisageables. Une procédure de pondération pourrait être de s'adapter aux spécificités de chaque strate au regard de la formule (17). L'estimation des paramètres nécessiterait alors d'utiliser un appariement entre les réponses passées à l'enquête de conjoncture et à une enquête quantitative. En particulier, les réponses aux évolutions passées et prévues de la production dans l'enquête sur l'activité dans l'industrie pourraient être comparées à celle des enquêtes de branches qui permettent le calcul de l'Indice de Production Industriel (IPI). Un tel appariement permettrait d'estimer strate par strate les différents paramètres (Q, σ_ϵ, σ_Z, s^+ et s^-). Les pondérations seraient alors construites à l'aide des poids relatifs λ et $1-\lambda$ à affecter à des pondéra-

[21] Rappelons que $\rho = \sigma_\epsilon / \sigma_Z$.

tions proportionnelles et constantes. Toutefois, cette procédure apparaît complexe à mettre en œuvre.

Dans cette étude, nous privilégions une procédure beaucoup plus simple qui consiste à tester des poids relatifs λ et $1-\lambda$ communs à l'ensemble des strates. Les pondérations peuvent ainsi être proportionnelles, constantes ou intermédiaires.

Deux types de pondérations intermédiaires sont envisagés ici. La première découle le plus naturellement possible du développement théorique précédent. Il consiste à choisir des poids qui soient des moyennes pondérées entre des poids proportionnels et des poids constants. Un calcul élémentaire montre que ceci est équivalent à faire *ex-post* une moyenne pondérée des soldes non pondérés et des soldes pondérés proportionnellement. La forme retenue peut s'écrire dans ce cas comme $p_i \propto q_i + \overline{q}$ où $\overline{q}$ est une constante et où $\propto$ désigne la proportionnalité. Le second type de pondérations intermédiaires envisagé consiste à pondérer selon une puissance k (inférieure à 1) appliquée à la taille de l'entreprise. On note alors $p_i \propto q_i^k$ avec $k \in]0,1[$. Les calculs empiriques montrent que cette deuxième procédure intermédiaire donne des résultats très proches de la première. Elle a de plus l'avantage de restreindre les conséquences d'une valeur aberrante parmi les mesures des tailles des entreprises. Dans le cas d'une taille d'entreprise $x_{i,-1}$ beaucoup trop grande à cause d'une erreur dans la déclaration de l'entreprise, cette erreur sera entièrement reflétée dans la première procédure (la constante devient alors négligeable). En revanche, avec cette seconde procédure, la puissance strictement inférieure à un limite l'effet de l'erreur sur les pondérations.

5 Applications empiriques

Plusieurs exemples sont présentés ci-après afin de comparer les qualités prédictives des soldes d'opinion en fonction des stratégies de pondérations primaires utilisées.

Le premier exemple porte sur les évolutions de l'emploi salarié dans l'industrie

manufacturière. Il utilise les questions relatives aux évolutions passées et prévues des effectifs dans l'enquête sur l'activité dans l'industrie. Cet exemple se place parfaitement dans les hypothèses du modèle développé précédemment. En particulier, les évolutions de l'agrégat macroéconomique sont faibles (σ_Z faible) et les développements limités du modèle sont donc valides. De plus, les strates peuvent être considérées comme homogènes puisque le découpage par strate est assez fin et prend en compte les tailles des entreprises.

Le second exemple porte sur les évolutions de la production manufacturière et utilise des questions relatives aux évolutions passées et prévues de la production dans l'enquête sur l'activité dans l'industrie. Les strates pour ces questions sont moins homogènes que pour celles relatives aux effectifs car il n'y a pas de stratification par taille dans le cas présent[22]. Ceci tend à favoriser des pondérations proportionnelles par rapport au cadre strict du modèle.

Le troisième exemple porte sur les évolutions de l'investissement des entreprises. Il utilise les questions relatives aux évolutions passées et prévues de l'investissement dans l'enquête sur les investissements dans l'industrie. Deux éléments font s'éloigner cet exemple du cadre du modèle. D'une part, le champ de l'enquête se limite à l'industrie alors que les évolutions à prévoir portent sur tous les secteurs de l'économie[23]. L'industrie ne réalise qu'un tiers des investissements de l'ensemble des entreprises. Ceci tend à favoriser des pondérations constantes par rapport au cadre strict du modèle puisque le lien entre les investissements de l'industrie et ceux des autres secteurs ne passe pas par une relation comptable mais simplement par une composante conjoncturelle commune. D'autre part, la variable agrégée est assez volatile, ce qui rend le développement limité du modèle moins légitime que dans les deux exemples précédents. De plus, remarquons que la forte hétérogénéité des comportements individuels en matière d'investissement (ρ fort) encourage des

[22] Les strates sont définies ici comme un regroupement de produits.

[23] Dans les comptes nationaux trimestriels, la Formation Brute de Capital Fixe (FBCF) des entreprises n'est pas détaillée par branche ou par secteur d'activité mais uniquement par produit.

pondérations proches de la proportionnalité, ce qui va à l'encontre de l'effet de différences de champs.

Dans tous les cas, quatre types de pondérations sont testés :

1. des pondérations constantes,
2. des pondérations en racine carrée,
3. des pondérations moyennes, calculées par $1/2(1/N_h + x_i/X_h)$, où N_h est le nombre de réponses dans la strate h de l'entreprise i, x_i la variable de taille pour l'entreprise i et X_h la somme des x_i dans la strate,
4. des pondérations proportionnelles à la taille.

De manière générale, avant de rentrer dans les cas particuliers, il apparaît que selon les cas, la meilleure stratégie consiste à prendre soit des pondérations constantes, soit des pondérations intermédiaires (pondérations en racine carrée ou pondérations moyennes), soit encore des pondérations proportionnelles à la taille de l'entreprise ou à celle de l'unité productive.

Il apparaît également que les deux stratégies intermédiaires (pondérations en racine carrée et pondérations moyennes) donnent des résultats toujours très proches.

5.1 Une application pratique concernant l'emploi et la production manufacturière dans l'enquête Activité dans l'industrie

Les données mobilisées dans cette partie sont issues des résultats trimestriels à l'enquête de l'Insee sur la situation et les perspectives dans l'industrie (désignée ci-après sous la dénomination raccourcie d'enquête Activité).

Les données individuelles issues de l'enquête Activité ont la particularité d'être de deux types : les questions « entreprises », pour lesquelles les entrepreneurs s'expriment sur l'évolution de caractéristiques de leur entreprise considérée globalement (questions sur l'évolution de l'emploi par exemple) et les questions de type

« produits » pour lesquelles les chefs d'entreprise répondent pour chacun des principaux produits déclarés[24] (questions sur l'évolution de la production par exemple).

Dans ce qui suit, on utilise des données relatives au secteur manufacturier tirées de l'enquête Activité. On considère les réponses données aux quatre questions suivantes :

– Évolution des effectifs au cours des trois derniers mois (TRE) : hausse / stabilité / baisse
– Évolution probable des effectifs au cours des trois prochains mois (TPE) : hausse / stabilité / baisse
– Évolution de la production au cours des trois derniers mois (TPPA) : hausse / stabilité / baisse
– Évolution probable de la production au cours des trois prochains mois (TPPRE) : hausse / stabilité / baisse.

Pour ces questions, les quatre stratégies de pondérations primaires ont été mises en œuvre. Les séries finalement obtenues après agrégation ont ensuite été désaisonnalisées (procédure X11-Arima).

Le tableau 1 (respectivement tableau 2) présente les corrélations entre les évolutions trimestrielles de l'emploi[25] (respectivement de la production[26]). Précisons que, du fait des périodes de référence visées par ces questions, les soldes des réponses à la question rétrospective (production passée) d'une enquête sont à rapprocher du taux de croissance trimestriel de la variable d'intérêt (production manufacturière) observée pour le trimestre précédant la date d'enquête.

Il apparaît que, pour les questions relatives aux effectifs, le choix des pondérations constantes apporte un gain en terme de corrélation, par rapport à la méthode uti-

[24] En moyenne, chaque entreprise déclare 1,4 produit.

[25] Évolution de l'emploi salarié de l'industrie manufacturière (à la fin du premier trimestre 2005).

[26] Mesurées par les comptes trimestriels - résultats détaillées du premier trimestre 2005 (base 2000).

TAB. 1 – Corrélations entre les soldes d'opinions et les évolutions de l'emploi dans l'industrie manufacturière

	Évolutions passées (TRE décalé *)	Évolutions prévues (TPE)
Pondérations constantes	93,4%	83,1%
Pondérations en racine carrée	92,8%	81,5%
Pondérations moyennes	93,0%	82,5%
Pondérations proportionnelles	91,9%	81,3%

Source : Insee, *Enquête activité dans l'industrie* et emploi salarié à la fin du premier trimestre 2005. Calculs des auteurs.

$*$: Décalé d'un trimestre de manière à faire correspondre le calendrier de la question à celui des comptes trimestriels

lisée actuellement (pondérations proportionnelles aux effectifs déclarés). Pour la question concernant l'évolution prévue des effectifs (TPE), la corrélation avec le taux de croissance de l'emploi manufacturier passe de 81,3% pour le solde calculé actuellement (pondération proportionnelle aux effectifs déclarés) à 83,1% pour celui calculé avec des pondération constantes. Les résultats sont similaires pour la question TRE.

Ceci illustre parfaitement les éléments théoriques développés dans cet article : en particulier l'hypothèse vérifiée ici pour les questions « entreprises », d'avoir des strates élémentaires (définies par le croisement taille $\times$ NES 114) contenant des unités très homogènes.

Au contraire, pour les questions « produits », les strates élémentaires correspondent à un regroupement de produits (NES 114). Les unités de cette strate, même si elles suivent une conjoncture commune (par exemple celle de l'« Industrie pharmaceutique (C31) »), ont néanmoins des caractéristiques propres (en terme d'organisation de leur production) qui dépendent effectivement de leur taille[27]. Dans ces conditions, la meilleure stratégie en terme de corrélation avec la production n'est pas,

[27] La distribution par taille des taux d'utilisation des capacités de production (TUC) fait apparaître par exemple des niveaux plus élevés pour les grandes entreprises que pour les autres, traduisant la facilité d'utilisation plus intense du capital lorsque les effectifs sont nombreux.

TAB. 2 – Corrélations entre les soldes d'opinions et les évolutions de la production manufacturière

	Évolutions passées (TPPA décalé *)	Évolutions prévues (TPPRE)
Pondérations constantes	63,6%	55,4%
Pondérations en racine carrée	67,8%	60,8%
Pondérations moyennes	66,8%	59,7%
Pondérations proportionnelles	68,9%	62,5%

Source : Insee, *Enquête activité dans l'industrie* et comptes trimestriels (résultats détaillés du premier trimestre de 2005). Calculs des auteurs.

$*$: Décalé d'un trimestre de manière à faire correspondre le calendrier de la question à celui des comptes trimestriels

comme précédemment, d'utiliser des pondérations constantes mais bien des pondérations proportionnelles à la taille de l'image de ce qui est fait actuellement pour le calcul des soldes publiés par l'Insee (pondérations égales aux chiffres d'affaires par produits déclarés par les entreprises elles-mêmes). On lit ainsi dans le tableau 2 que la corrélation de la variable de production prévue (TPPRE) avec la variable d'intérêt (taux de croissance de la production manufacturière du trimestre courant) est maximale à 62,5% dans le cas des soldes avec pondérations proportionnelles. Cette corrélation diminue ensuite lorsqu'on adopte une stratégie de pondération alternative : 55,4% (pondérations constantes) ou 60,8% (pondération en racine carrée). Les résultats pour la variable TPPA vont dans le même sens, même si la perte en terme de corrélation est un peu moins nette lorsqu'on ne choisit pas les pondérations proportionnelles.

5.2 Une application pratique concernant les évolutions semestrielles de l'investissement et l'enquête sur les investissements dans l'industrie

L'enquête sur les investissements dans l'industrie est trimestrielle. Cette enquête est relativement atypique pour une enquête de conjoncture puisque ses principales questions sont quantitatives. Toutefois, l'enquête présente également deux ques-

tions qualitatives à trois modalités. Ces questions portent sur les évolutions semestrielles courantes et prévues des investissements.

En avril et en octobre, les industriels sont ainsi interrogés sur le niveau de leurs investissements durant le semestre courant par rapport au semestre précédent et sur celui de leurs investissements prévus pour le prochain semestre par rapport au semestre courant.

Ces deux questions permettent de construire un solde relatif aux évolutions courantes (SEMPA) et un solde relatif aux évolutions anticipées (SEMPR). Les deux soldes sont alors naturellement mis en regard des évolutions de l'investissement. Plus précisément, ils sont comparés aux évolutions en volume et en moyenne semestrielle de la FBCF[28] des Entreprises Non Financières (ENF). Le solde SEMPA est comparé aux évolutions courantes. Le solde SEMPR est naturellement comparé aux évolutions du semestre suivant. Toutefois, il apparaît que le solde SEMPR est mieux corrélé aux évolutions courantes et on le compare donc également à ces évolutions concomitantes.

Le champ de l'enquête ne porte que sur l'industrie. Il est donc très restrictif par rapport à l'ensemble des ENF. Le tableau 3 présente les corrélations entre les évolutions semestrielles (mesurées par les comptes trimestriels avec les résultats détaillés du premier trimestre de 2005) et les différents soldes.

Concernant les évolutions courantes de l'investissement (SEMPA), la meilleure stratégie de pondération apparaît d'utiliser les effectifs avec des pondérations intermédiaires entre des pondérations constantes et des pondérations proportionnelles aux effectifs. Les deux exemples des pondérations en racine carrée et des pondérations moyennes fournissent des soldes très corrélés aux évolutions de la FBCF des ENF (75%). La méthode retenue pour les soldes publiés par l'Insee consiste à

[28] Formation Brute de Capital Fixe.

TAB. 3 – Corrélations entre les soldes d'opinions et les évolutions de l'investissement des Entreprises Non Financières (ENF)

	SEMPA	SEMPR	SEMPR décalé *
Pondérations constantes			
	72,2%	58,7%	53,0%
Pondérations en racine carrée			
CA	72,6%	60,8%	46,3%
Investissement	68,9%	58,6%	42,9%
Effectifs	75,2%	59,5%	46,8%
Pondérations moyennes			
CA	72,3%	60,7%	45,0%
Investissement	68,8%	60,1%	42,0%
Effectifs	75,0%	58,4%	43,7%
Pondérations proportionnelles			
CA	68,9%	54,4%	35,1%
Investissement	62,1%	53,2%	30,3%
Effectifs	74,2%	51,3%	32,1%

Source : Insee, *Enquête investissement* et comptes trimestriels (résultats détaillés du premier trimestre de 2005). Calculs des auteurs.

$*$: Décalé d'un semestre de manière à faire correspondre le calendrier de la question à celui des comptes trimestriels

pondérer proportionnellement aux investissements. Avec une telle stratégie, la corrélation recule à 62,1% (cf. la Fig. 9 dans l'annexe C).

De ce point de vue, les investissements n'apparaissent pas être une variable de pondération optimale (cf. la Fig. 10 dans l'annexe C). La présence de montants d'investissements nuls conduit à ne pas prendre en compte un grand nombre de réponses. Plus généralement, l'investissement est une donnée volatile et les investissements d'une année donnée n'indiquent pas nécessairement bien l'importance des investissements de l'entreprise sur une période plus longue (cf. la Fig. 10 dans l'annexe C).

Quelle que soit la variable retenue (chiffre d'affaires, montant d'investissement ou effectifs), les stratégies de pondérations qui apparaissent comme optimales sont les deux méthodes intermédiaires entre des pondérations proportionnelles et des pondérations constantes (cf. la Fig. 11 de l'annexe C).

Concernant les soldes relatifs aux évolutions prévues de l'investissement (SEMPR), il est à remarquer tout d'abord que les soldes SEMPR sont davantage corrélés aux évolutions concomitantes qu'aux évolutions futures. Si on se réfère aux évolutions concomitantes, les meilleures stratégies de pondération apparaissent être également les stratégies intermédiaires. Au contraire, si on se réfère aux évolutions du semestre suivant (SEMPR décalé), les soldes non pondérés sont les plus corrélés aux évolutions. Dans les deux cas, la variable à utiliser semble être en priorité les chiffres d'affaires, suivie de près par les effectifs. Comme pour les soldes passés, les montants d'investissement apparaissent moins pertinents comme variable de pondération.

TAB. 4 – Résumé des résultats empiriques selon les questions étudiées

	Pondérations constantes	Pondérations intermédiaires	Pondérations proportionnelles
Emploi dans l'industrie manufacturière	·|· ⌣	·|· =	·|· ⌢
Production manufacturière	·|· ⌢	·|· =	·|· ⌣
Investissement des entreprises	·|· =	·|· ⌣	·|· ⌢

5.3 Empiriquement, le choix des pondérations primaires dépend de la stratification et de la variable étudiée

Les tests des différents types de pondérations envisagés pour les six questions détaillées ci-dessus permettent de retrouver de manière empirique ce que les raisonnements théoriques laissaient présager :

Emploi (TRE et TPE) On est dans le cadre exact du modèle. Les pondérations optimales semblent plus proches des pondérations constantes que des pondérations intermédiaires proposées ici.

Production manufacturière (TPPA et TPPRE) Il y a beaucoup d'hétérogénéité des strates du fait de l'absence de stratification par taille d'entreprises. Les pondérations optimales semblent plus proches de pondérations proportionnelles.

Investissement (SEMPA et SEMPR) Dans ce cas, les meilleurs pondérations apparaissent être des pondérations intermédiaires. Deux phénomènes se compensent. D'une part, les investissements ont une forte composante idiosyncrasique. Ceci encourage à utiliser des pondérations proches de la proportionnalité. D'autre part, l'enquête ne couvre pas tous les secteurs mais uniquement celui de l'industrie. Le caractère *out of sample* s'en trouve renforcé, ce qui encourage des pondérations proches de pondérations constantes.

Il apparaît donc que le choix des pondérations primaires dépend de la question. Le modèle développé permet d'éclairer le choix des pondérations primaires. Toutefois,

un examen *a posteriori* des corrélations peut compléter efficacement le critère de choix des pondérations primaires.

6 Conclusion

Le modèle présenté ici est d'une expression plus développée que celle présentée par Fansten (1976). Fansten concluait sur le fait que les pondérations primaires étaient neutres dans l'introduction du solde d'opinion pour résumer l'information rassemblée par une question qualitative. Au contraire, il apparaît ici que le solde d'opinion ne peut résumer au mieux l'information contenue dans les réponses que sous couvert que les pondérations soient correctement choisies.

Le modèle théorique propose comme pondérations primaires une fonction affine et croissante de la taille des variables individuelles considérées. Autrement dit les pondérations optimales apparaissent comme des pondérations intermédiaires entre des pondérations constantes (cas non pondéré) et des pondérations proportionnelles à la taille des entreprises.

Ces conclusions sont vérifiées empiriquement sur quelques questions dans les enquêtes de conjoncture menées par l'Insee dans l'industrie. En fonction des cas étudiés, il apparaît bien que les pondérations proportionnelles ne sont pas nécessairement les meilleures. Les pondérations primaires peuvent alors être choisies en fonction de l'existence d'une stratification par taille pour l'enquête, de son taux de couverture et de la volatilité de la variable considérée.

Cette étude pourrait sans doute être poursuivie dans deux directions : d'une part, l'étude empirique pourrait être généralisée à d'autres questions quantitatives et à d'autres enquêtes de conjoncture, en particulier dans d'autres secteurs d'activité (commerce, services, ...). D'autre part, un croisement des enquêtes de conjoncture avec des enquêtes quantitatives mesurant *a posteriori* les variations individuelles réelles permettraient d'enrichir l'analyse empirique et de mettre en place des mé-

thodes de choix des pondérations en fonction des caractéristiques propres à chaque strate.

Références

[1] Fiche méthodologique : Enquête sur les investissements dans l'industrie, sur le site www.insee.fr. Sous la rubrique conjoncture/indicateurs de conjoncture/principaux indicateurs.

[2] Fiche méthodologique : Enquête mensuelle de conjoncture dans l'industrie, sur le site www.insee.fr. Sous la rubrique conjoncture/indicateurs de conjoncture/principaux indicateurs.

[3] Fiche méthodologique : Enquête trimestrielle de conjoncture dans l'industrie, sur le site www.insee.fr. Sous la rubrique conjoncture/indicateurs de conjoncture/principaux indicateurs.

[4] **Anderson O. (1951)**, Konjoncturtest und Statistik, *Allgemeines Statistisches Archiv, vol. 35, p. 209-220.*

[5] **Anderson O. (1952)**, The business test of the IFO-Institute for Economic Research, Munich, and its theoretical model, *Review of International Statistical Institute, 20, p. 1-17.*

[6] **Biau O., H. Erkel-Rousse et N. Ferrari (2005)**, Réponses individuelles aux enquêtes de conjoncture et prévision macroéconomique : Exemple de la prévision de la production manufacturière, *Insee, document de travail n°G2005/12.*

[7] **Batchelor R. (1981)**, Aggregate expectation under the stable laws, *Journal of Econometrics, 16, p. 199-210.*

[8] **Carlson J.A. et M. Parkin (1975)**, Inflation expectations, *Economica, 42, p. 123-138.*

[9] **Cunningham a.W.F., r.J. Smith et M.R. Weale (1998)**, Measurement errors and data estimation : the quantification of survey data, *Applied Economics and Public Policy, Cambridge University Press.*

[10] **D'Elia E. (2005)**, Using the results of qualitative surveys in quantitative analysis, *ISAE, working paper n°56.*

[11] **Doz C., F. Lenglart (1999)**, Analyse factorielle dynamique : test du nombre de facteurs, estimation et application à l'enquête de conjoncture dans l'industrie, *Annales d'Économie et Statistique, n°54, p. 91-127.*

[12] **Fansten M. (1976)**, Introduction à une théorie mathématique de l'opinion, *Annales de l'Insee, n° 21 - 1976.*

[13] **Fayolle J. (1987)**, Pratique contemporaine de l'analyse conjoncturelle, *Economica.*

[14] **Gregoir S. et F. Lenglart (1998)**, Un nouvel indicateur pour saisir les retournements de conjoncture, *Économie et Statistique, n°314, p. 39-60.*

[15] **Hild F. (2002)**, Une lecture enrichie des réponses aux enquêtes de conjoncture, *Économie et statistique, n° 359-360 - 2002.*

[16] **Mitchell J., R.J. Smith et M.R. Weale (2002)**, Quantification of qualitative firm-level survey data, *Economic journal, à paraître.*

[17] **Mitchell J. (2002)**, The use of non-normal distributions in quantifying qualitative survey data on expectations, *Economics Letters, à paraître.*

[18] **Pesaran M.H. (1985)**, Formation of Inflation Expectations in British Manufacturing Industries, *The Economic Journal, 95, p. 948-975.*

[19] **Pesaran M.H. (1984)**, Expectations formation and macroeconomic modelling, *Contemporary Macroeconomic Modelling (ed. P. Malgrange and P.A. Muet) Oxford : Blackwell.*

[20] **Ronning G. (1985)**, Econometric approaches to the estimation of indifference intervals in business tendency surveys, *17th CIRET Conference, Vienne.*

[21] **Theil H. (1952)**, On the time shape of economic microvariables and the Munich business test, *Revue de l'Institut International de Statistique, 20.*

[22] **Theil H. (1955)**, Recent Experiences with the Munich Business Test : An Expository Article, *Econometrica, vol. 23, n°2, p. 184-192.*

[23] **Theil H. et J.S. Cramer (1954)**, On the Utilisation of a New Source of Economic Information, *Paper read at the Uppsala Meeting of the Econometric Society*.

A Modèle multiplicatif et distributions log-normales

Dans le corps du texte, nous avons développé un modèle sous forme additive : les variations individuelles $T_i = X_i/x_{i,-1} - 1$ sont la somme d'une composante commune Z et d'une composante individuelles ϵ_i. Cette approche a l'avantage de la simplicité. Toutefois, elle apparaît moins réaliste qu'un modèle multiplicatif. Dans ce cas, on écrit les évolutions individuelles comme $T_i = Z\epsilon_i$. Les lois suivies par les ϵ_i et par Z sont alors de lois log-normales. Cette approche est en particulier celle qui est retenue dans l'article de référence sur la théorie de l'opinion de Fansten (1976).

Plus précisément, dans le modèle additif développé dans le corps du texte, la symétrie des distributions est nécessaire à l'introduction des soldes d'opinion. Selon l'hypothèse naturelle de taux de croissance suivant des distributions log-normales, la symétrie des lois est alors à rejeter.

Afin de montrer que les conclusions du corps du texte ne sont pas rejetées dans le cas d'une modélisation multiplicative, nous développons ici un tel modèle avec des lois log-normales. Toutefois, cette approche s'avère plus difficile à écrire et nous nous ne développerons ici que l'équivalent de la partie 3.1, c'est-à-dire avec l'hypothèse simplificatrice portant sur les prévisions *in sample* : pour la prévision des variations T_i d'une entreprise i de l'échantillon, on considéra que les réponses des autres entreprises j ($j \in \mathcal{E} \setminus \{i\}$) de l'échantillon n'apportent pas d'information supplémentaire à celle donnée par la réponse s_i de l'entreprise i.

A.1 Définitions et rappels pour les lois log-normales

Soit Y une variable aléatoire positive telle que $\log(Y)$ suit une loi normale. On dit alors que Y suit une loi log-normale. On notera m_Y et $m_{\log(Y)}$ les moyennes de Y et de $\log(Y)$ et σ_Y^2 et $\sigma_{\log(Y)}^2$ leur variance. On notera également f_Y (respectivement F_Y) et $f_{\log Y}$ (respectivement $F_{\log Y}$) les densités (respectivement les fonctions de répartitions) des loi suivies par Y et $\log Y$. On note enfin f_1 (respecti-

vement F_1) la densité (respectivement la fonction de répartition) de la loi normale centrée et de variance unitaire.

Plusieurs résultats seront nécessaires par la suite. Tout d'abord, on a :

$$\left\{ \begin{array}{rcl} F_Y(y) & = & F_{\log Y}(\log y) = F_1\left(\frac{\log y - m_{\log Y}}{\sigma_{\log Y}}\right) \\ f_Y(y) & = & \frac{1}{y} f_{\log Y}(\log y) = \frac{1}{y}\frac{1}{\sigma_{\log Y}} f_1\left(\frac{\log y - m_{\log Y}}{\sigma_{\log Y}}\right) \end{array} \right.$$

L'espérance et la variance de Y sont données par :

$$\left\{ \begin{array}{rcll} m_Y & = & E[Y] & = \exp\left(m_{\log Y} + \frac{\sigma^2_{\log Y}}{2}\right) \\ \sigma^2_Y & = & E[Y^2] - E[Y]^2 & = \exp\left(m_{\log Y} + \frac{\sigma^2_{\log Y}}{2}\right)^2 \left(\exp\left(\sigma^2_{\log Y}\right) - 1\right) \end{array} \right.$$

Enfin, dans la partie A.3, on utilisera :

$$\begin{array}{rl} \int_a^b Y\, f_Y(y)dy & = \int_{\log a}^{\log b} e^{\xi}\, f_{\log Y}(\xi)d\xi \\ & = F_1\left(\frac{\log b - m_{\log Y}}{\sigma_{\log Y}} - \sigma_{\log Y}\right) - F_1\left(\frac{\log a - m_{\log Y}}{\sigma_{\log Y}} - \sigma_{\log Y}\right) \end{array}$$

A.2 Pondérations pour les prévisions *out of sample*

Les Z sont i.i.d. et leurs logarithmes sont distribués selon la loi normale de centre $m_{\log Z}$ et de variance $\sigma^2_{\log Z}$. Les ϵ_i sont i.i.d. et leurs logarithmes sont distribués selon la loi normale de centre $m_{\log \epsilon} = 0$ et de variance $\sigma^2_{\log \epsilon}$. Les Z et les ϵ_i sont indépendants entre eux.

Nous nous intéressons tout d'abord aux prévisions des variations *out of sample*. Nous écrivons :

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}} \,|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = E\left[Z\epsilon_i \,|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}\right] \text{ avec } i \in \mathcal{P}\setminus\mathcal{E} \qquad (18)$$

$$E\left[Z\epsilon_i \,|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}\right] = E_Z\left[ZE_\epsilon\epsilon_i \,|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}\right] \text{ avec } i \in \mathcal{P}\setminus\mathcal{E}$$

$$E\left[Z\epsilon_i \,|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}\right] = m_\epsilon E_Z\left[Z \,|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}\right]$$

On rappelle que $m_{\log\epsilon} = 0$ et donc que $m_\epsilon = \exp\left(\frac{\sigma^2_{\log\epsilon}}{2}\right)$. D'autre part, la formule de Bayes (3) reste vraie et la formule (4) se récrit :

$$\begin{aligned} P\left(\{S_i\}_{i\in\mathcal{E}} = \{s_i\}_{i\in\mathcal{E}} | Z = z\right) &= \left[1 - F_\epsilon(s^+/z)\right]^{N^+} \\ &\times \left[F_\epsilon(s^+/z) - F_\epsilon(s^-/z)\right]^{N^0} \\ &\times F_\epsilon(s^-/z)^{N^-} \end{aligned} \tag{19}$$

Soit encore :

$$\begin{aligned} P\left(\{S_i\}_{i\in\mathcal{E}} = \{s_i\}_{i\in\mathcal{E}} | Z = z\right) &= \left[1 - F_{\log\epsilon}(\log(s^+/z))\right]^{N^+} \\ &\times \left[F_{\log\epsilon}(\log(s^+/z)) - F_{\log\epsilon}(\log(s^-/z))\right]^{N^0} \\ &\times F_{\log\epsilon}(\log(s^-/z))^{N^-} \end{aligned} \tag{20}$$

Pour simplifier l'écriture, on note :

$$\left\{\begin{array}{rcl} \mu &=& \exp\left(m_{\log Z}\right) \\ F_+ &=& F_{\log\epsilon}\left(\log(s^+/\mu)\right) \\ F_- &=& F_{\log\epsilon}\left(\log(s^-/\mu)\right) \\ f_+ &=& f_{\log\epsilon}\left(\log(s^+/\mu)\right) = F'_{\log\epsilon}\left(\log(s^+/\mu)\right) \\ f_- &=& f_{\log\epsilon}\left(\log(s^-/\mu)\right) = F'_{\log\epsilon}\left(\log(s^-/\mu)\right) \\ f'_+ &=& f'_{\log\epsilon}\left(\log(s^+/\mu)\right) \\ f'_- &=& f'_{\log\epsilon}\left(\log(s^-/\mu)\right) \end{array}\right.$$

$$\begin{aligned} P\left(\{S_i\}_{i\in\mathcal{E}} = \{s_i\}_{i\in\mathcal{E}} | Z = z\right) &= (1 - F_+)^{N^+} \times (F_+ - F_-)^{N^0} \times F_-^{N^-} \\ &\times \left[1 + N\tilde{\Gamma}\log(z/\mu) + \left(N\tilde{\Phi} + N^2\tilde{\Omega}\right)\log(z/\mu)^2\right] \\ &+ o\left(\log(z/\mu)^2\right) \end{aligned}$$

avec :

$$\begin{cases} \tilde{\Gamma} &= \frac{f_+}{1-F_+}\tilde{R}^+ - \frac{f_+-f_-}{F_+-F_-}\tilde{R}^0 - \frac{f_-}{F_-}\tilde{R}^- \\ \tilde{\Phi} &= \frac{1}{2}\tilde{R}^-\left(\frac{f'_-}{F_-} - \frac{f_-^2}{F_-^2}\right) + \frac{1}{2}\tilde{R}^0\left(\frac{f'_+-f'_-}{F_+-F_-} - \frac{(f_+-f_-)^2}{(F_+-F_-)^2}\right) + \frac{1}{2}\tilde{R}^+\left(-\frac{f'_+}{1-F_+} - \frac{f_+^2}{(1-F_+)^2}\right) \\ \tilde{\Omega} &= \tilde{R}^-\tilde{R}^0\frac{f_-}{F_-}\frac{f_+-f_-}{F_+-F_-} - \tilde{R}^0\tilde{R}^+\frac{f_+-f_-}{F_+-F_-}\frac{f_+}{1-F_+} - \tilde{R}^+\tilde{R}^-\frac{f_+}{1-F_+}\frac{f_-}{F_-} \\ &\quad +\frac{1}{2}\left(\tilde{R}^-\right)^2\frac{f_-^2}{F_-^2} + \frac{1}{2}\left(\tilde{R}^0\right)^2\frac{(f_+-f_-)^2}{(F_+-F_-)^2} + \frac{1}{2}\left(\tilde{R}^+\right)^2\frac{f_+^2}{1-F_+^2} \end{cases}$$

Afin d'alléger l'écriture, on note également $\alpha_1 = N\tilde{\Gamma}$ et $\alpha_2 = N\tilde{\Phi} + N^2\tilde{\Omega}$. La formule de Bayes donne :

$$E\left[Z|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = \frac{\int_{-\infty}^{+\infty} z\cdot[1+\alpha_1\log(z/\mu)+\alpha_2\log(z/\mu)^2+o(\log(z/\mu)^2)]\cdot f_Z(z)\,dz}{\int_{-\infty}^{+\infty}[1+\alpha_1\log(z/\mu)+\alpha_2\log(z/\mu)^2+o(\log(z/\mu)^2)]\cdot f_Z(z)\,dz}$$

En faisant le changement de variable $\log z = \zeta$ et en utilisant un développement limité de $\exp(\zeta)$ en $\zeta = m_{\log Z}$, on obtient :

$$E\left[Z|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] =$$

$$\mu\frac{\int_{-\infty}^{+\infty}[1+(1+\alpha_1)(\zeta-m_{\log Z})+(\frac{1}{2}+\alpha_1+\alpha_2)(\zeta-m_{\log Z})^2+o((\zeta-m_{\log Z})^2)]\cdot f_{\log Z}(\zeta)\,d\zeta}{\int_{-\infty}^{+\infty}[1+\alpha_1(\zeta-m_{\log Z})+\alpha_2(\zeta-m_{\log Z})^2+o((\zeta-m_{\log Z})^2)]\cdot f_{\log Z}(\zeta)\,d\zeta}$$

$$E\left[Z|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = \mu\left[1+\left(\frac{1}{2}+\alpha_1\right)\sigma^2_{\log Z}\right] + o\left(\sigma^2_{\log Z}\right)$$

Soit encore :

$$E\left[Z|\mathcal{I}_{-1}, \{s_i\}_{i\in\mathcal{E}}\right] = m_Z\left(1+N\tilde{\Gamma}\sigma^2_{\log Z}\right) + o\left(\sigma^2_{\log Z}\right)$$

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]=m_Z\exp\left(\frac{\sigma^2_{\log\epsilon}}{2}\right)\left(1+N\tilde{\Gamma}\sigma^2_{\log Z}\right)+o\left(\sigma^2_{\log Z}\right)\quad(21)$$

Ceci se récrit encore, avec $m_{\log T}$ et $\sigma^2_{\log T}$ la moyenne et la variance de la distribution des logarithmes des T_i ($m_{\log T}=m_{\log Z}$ et $\sigma^2_{\log T}=\sigma^2_{\log Z}+\sigma^2_{\log\epsilon}$) :

$$E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]=m_T\left(1+N\tilde{\Gamma}\sigma^2_{\log Z}\right)+o\left(\sigma^2_{\log Z}\right)\qquad(22)$$

Sous une forme adaptée à la présentation en modèle multiplicatif, on retiendra le résultat donné par la formule (23).

$$\log\left(E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]\right)=\log\left(m_T\right)+N\tilde{\Gamma}\sigma^2_{\log Z}+o\left(\sigma^2_{\log Z}\right)\qquad(23)$$

De la même manière que dans le modèle additif, les proportions de réponses non-pondérées (par l'intermédiaire de $\tilde{\Gamma}$) résument entièrement l'information fournie par les réponses des entreprises de l'échantillon pour la prévision *out of sample*.

On veut ensuite introduire le solde d'opinion non-pondéré $\tilde{S}$. Comme dans le modèle additif, une hypothèse de symétrie des seuils psychologiques est nécessaire. Cette condition de symétrie s'écrit maintenant $\mu=\sqrt{s^+s^-}$. Sous cette condition, on peut alors définir $\delta_{\log\epsilon}$ par $\delta_{\log\epsilon}=\frac{\log(s^+/\mu)}{\sigma_{\log\epsilon}}$ ou par $\delta_{\log\epsilon}=-\frac{\log(s^-/\mu)}{\sigma_{\log\epsilon}}$. On définit f_1 et F_1 les fonctions de densité et de répartition de la loi normale de moyenne nulle et de variance unitaire. On note également η_1 la fonction $f_1/(1-F_1)$ et ρ le rapport $\sigma_{\log\epsilon}/\sigma_{\log Z}$. On a alors :

$$\log\left(E\left[T^{\mathcal{P}\setminus\mathcal{E}}\,|\mathcal{I}_{-1},\{s_i\}_{i\in\mathcal{E}}\right]\right)=\log\left(m_T\right)+\sigma_{\log Z}N\frac{1}{\rho}\,\eta_1(\delta_{\log\epsilon})\tilde{S}+o\left(\sigma^2_{\log Z}\right)\qquad(24)$$

A.3 Pondérations pour les prévisions *in sample*

On cherche à calculer l'espérance de T_i pour l'entreprise i de l'échantillon connaissant toutes les réponses des entreprises de l'échantillon. Il serait possible d'utiliser de nouveau la modélisation $T_i = Z\epsilon_i$. Cette approche correspondrait alors à la partie 4.1 du modèle additif. Toutefois, les calculs sont alors particulièrement compliqués et il n'est pas possible d'introduire les soldes. Nous nous limitons donc ici à l'approche de la partie 3.1 qui consiste à négliger l'information apportée par les autres entreprises de l'échantillon $\mathcal{E}$ sur la composante conjoncturelle commune Z. Ceci revient à écrire :

$$E\left[T_i|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}, s_i = s\right] = E\left[T_i|\mathcal{I}_{-1}, s_i = s\right]$$

T_i suit une loi log-normale telle que $\log(T_i)$ soit de moyenne $m_{\log T}$ et de variance $\sigma^2_{\log T}$. Avec les notations de la partie A.2, on a $m_{\log T} = m_{\log Z}$ et $\sigma^2_T = \sigma^2_{\log Z} + \sigma^2_{\log \epsilon}$. On note de plus $\rho = \sigma_{\log \epsilon}/\sigma_{\log Z}$, d'où $\sigma_{\log T} = \sigma_{\log Z}\rho\sqrt{1+\frac{1}{\rho^2}}$.

$$\left\{\begin{array}{rcl} E[T_i|s_i = -1] &=& \frac{\int_0^{s^-} T\, f_T(T)\, dT}{\int_0^{s^-} f_T(T)\, dT} \\ E[T_i|s_i = \ \ 0] &=& \frac{\int_{s^-}^{s^+} T\, f_T(T)\, dT}{\int_{s^-}^{s^+} f_T(T)\, dT} \\ E[T_i|s_i = \ \ 1] &=& \frac{\int_{s^+}^{+\infty} T\, f_T(T)\, dT}{\int_{s^+}^{+\infty} f_T(T)\, dT} \end{array}\right.$$

Pour introduire les soldes d'opinion, on suppose également $\mu = \sqrt{s^+ s^-}$. On pose $\delta_{\log T} = \log(s^+/\mu)/\sigma_{\log T} = -\log(s^-/\mu)/\sigma_{\log T}$. On pose enfin $\eta_1(\delta_{\log T}) = f_1(\delta_{\log T})/(1 - F_1(\delta_{\log T}))$.

$$\left\{\begin{array}{rcl} E[T_i|s_i = -1] &=& m_T \frac{F_1(-\delta_{\log T} - \sigma_{\log T})}{F_1(-\delta_{\log T})} \\ E[T_i|s_i = \ \ 0] &=& m_T \frac{F_1(\delta_{\log T} - \sigma_{\log T}) - F_1(-\delta_{\log T} - \sigma_{\log T})}{F_1(\delta_{\log T}) - F_1(-\delta_{\log T})} \\ E[T_i|s_i = \ \ 1] &=& m_T \frac{F_1(\delta_{\log T} - \sigma_{\log T})}{F_1(\delta_{\log T})} \end{array}\right.$$

$$\left\{\begin{array}{rcl} E[T_i|s_i=-1] &=& m_T\left(1-\sigma_{\log T}\eta_1(\delta_{\log T})+\frac{1}{2}\sigma^2_{\log T}\frac{f_1'(-\delta_{\log T})}{F_1(-\delta_{\log T})}\right)+o\left(\sigma^2_{\log T}\right)\\ E[T_i|s_i=\ \ 0] &=& m_T\left(1+\sigma^2_{\log T}\frac{f_1'(-\delta_{\log T})}{F_1(-\delta_{\log T})}\right)+o\left(\sigma^2_{\log T}\right)\\ E[T_i|s_i=\ \ 1] &=& m_T\left(1+\sigma_{\log T}\eta_1(\delta_{\log T})+\frac{1}{2}\sigma^2_{\log T}\frac{f_1'(-\delta_{\log T})}{F_1(-\delta_{\log T})}\right)+o\left(\sigma^2_{\log T}\right)\end{array}\right.$$

$$E[T^{\mathcal{E}}|\{s_i\}_i]=m_T\left(1+\sigma_{\log T}\eta_1(\delta_{\log T})\hat{S}\right)+o\left(\sigma_{\log T}\right)$$

A.4 Pondérations pour les prévisions de la population toute entière

L'agrégation des prévisions *in sample* et *out of sample* donne alors la formule ci-dessous :

$$E[T|\{s_i\}_i]\approx m_T\left(1+Q\sigma_{\log Z}\rho\sqrt{1+\frac{1}{\rho^2}}\eta_1(\delta_{\log T})\hat{S}+(1-Q)\sigma_{\log Z}N\frac{1}{\rho}\,\eta_1(\delta_{\log\epsilon})\tilde{S}\right)$$

Les poids relatifs à donner aux soldes non pondérés et aux soldes pondérés proportionnellement à la taille des entreprises sont alors donnés respectivement par $1-\lambda$ et λ avec :

$$\frac{\lambda}{1-\lambda}=\frac{Q}{1-Q}\frac{\frac{\eta_1(\delta_T)}{\eta_1(\delta_\epsilon)}\sqrt{1+1/\rho^2}\,\rho^2}{N} \tag{25}$$

Avec $\sigma_Z \ll \sigma_\epsilon$ on a $\eta_1(\delta_\epsilon)\approx\eta_1(\delta_T)$ et $\rho\gg 1$. D'où :

$$\frac{\lambda}{1-\lambda}=\frac{Q}{1-Q}\frac{\rho^2}{N} \tag{26}$$

On retrouve ainsi un résultat de la forme de celui de la partie 3.3 dans le modèle additif. On rappelle que ce résultat correspond à l'hypothèse simplificatrice consistant à négliger l'information apportée par la composante conjoncturelle commune Z lors de la prévision *in sample*. Comme nous l'avons vu dans la partie 4.1 du modèle additif, ceci revient à surestimer l'importance donnée aux soldes pondérés proportionnellement à la taille des entreprises.

B Calculs pour la partie 4.1

$$\left\{\begin{array}{rcl} f(\epsilon|Z=z,s_i=-1) & = & \frac{\mathbf{1}_{\{\epsilon<s^- -z\}}\cdot f(\epsilon)}{P(s_i=-1|Z=z)} \\ f(\epsilon|Z=z,s_i=\ \ 0) & = & \frac{\mathbf{1}_{\{s^- -z\leq\epsilon<s^+ -z\}}\cdot f(\epsilon)}{P(s_i=\ \ 0|Z=z)} \\ f(\epsilon|Z=z,s_i=\ \ 1) & = & \frac{\mathbf{1}_{\{s^+ -z\leq\epsilon\}}\cdot f(\epsilon)}{P(s_i=\ \ 1|Z=z)} \end{array}\right.$$

$$\left\{\begin{array}{rcl} E_\epsilon[\epsilon_i|Z=z,s_i=-1] & = & \frac{\int_{-\infty}^{s^- -z}\epsilon\ f_\epsilon(\epsilon)\,d\epsilon}{\int_{-\infty}^{s^- -z} f_\epsilon(\epsilon)\,d\epsilon} \\ E_\epsilon[\epsilon_i|Z=z,s_i=\ \ 0] & = & \frac{\int_{s^- -z}^{s^+ -z}\epsilon\ f_\epsilon(\epsilon)\,d\epsilon}{\int_{s^- -z}^{s^+ -z} f_\epsilon(\epsilon)\,d\epsilon} \\ E_\epsilon[\epsilon_i|Z=z,s_i=\ \ 1] & = & \frac{\int_{-\infty}^{s^+ -z}\epsilon\ f_\epsilon(\epsilon)\,d\epsilon}{\int_{-\infty}^{s^+ -z} f_\epsilon(\epsilon)\,d\epsilon} \end{array}\right.$$

On note :

$$\left\{\begin{array}{l} I_- = \int_{-\infty}^{s^- -m_Z}\epsilon\ f_\epsilon(\epsilon)\,d\epsilon \\ I_0 = \int_{s^- -m_Z}^{s^+ -m_Z}\epsilon\ f_\epsilon(\epsilon)\,d\epsilon \\ I_+ = \int_{s^+ -m_Z}^{+\infty}\epsilon\ f_\epsilon(\epsilon)\,d\epsilon \end{array}\right. , \left\{\begin{array}{l} F_- = \int_{-\infty}^{s^- -m_Z} f_\epsilon(\epsilon)\,d\epsilon \\ F_+ = \int_{-\infty}^{s^+ -m_Z} f_\epsilon(\epsilon)\,d\epsilon \end{array}\right. \text{ et } \left\{\begin{array}{l} f_- = f_\epsilon(s^- - m_Z) \\ f_+ = f_\epsilon(s^+ - m_Z) \end{array}\right.$$

$$\left\{\begin{array}{rcl} E_\epsilon[\epsilon_i|Z=z,s_i=-1] & = & C_-\left\{1+D_-(z-m_Z)+G_-(z-m_Z)^2\right\}+o((z-m_Z)^2) \\ E_\epsilon[\epsilon_i|Z=z,s_i=\ \ 0] & = & C_0\left\{1+D_0(z-m_Z)+G_0(z-m_Z)^2\right\}+o((z-m_Z)^2) \\ E_\epsilon[\epsilon_i|Z=z,s_i=\ \ 1] & = & C_+\left\{1+D_+(z-m_Z)+G_+(z-m_Z)^2\right\}+o((z-m_Z)^2) \end{array}\right.$$

avec :

$$(a)\ \left\{\begin{array}{l} C_- = \frac{I_-}{F_-} \\ C_0 = \frac{I_0}{F_+ - F_-} \\ C_+ = \frac{I_+}{1-F_+} \end{array}\right.$$

$$(b)\ \left\{\begin{array}{l} D_- = \frac{f_-}{F_-} - \frac{(s^- - m_Z)f_-}{I_-} \\ D_0 = \frac{f_+ - f_-}{F_+ - F_-} - \frac{(s^+ - m_Z)f_+ - (s^- - m_Z)f_-}{I_0} \\ D_+ = -\frac{f_+}{1-F_+} + \frac{(s^+ - m_Z)f_+}{I_+} \end{array}\right.$$

$$(c)\ \begin{cases} G_- = & \frac{1}{2}\frac{(s^- - m_Z)f'_- + f_-}{I_-} - \frac{(s^- - m_Z)f_-^2}{I_-\,F_-} - \frac{1}{2}\frac{f'_-}{F_-} + \frac{f_-^2}{F_-^2} \\ G_0 = & \frac{1}{2}\frac{(s^+ - m_Z)f'_+ + f_+ - (s^- - m_z)f'_- - f_-}{I_0} - \frac{\big((s^+ - m_Z)f_+ - (s^- - m_Z)f_-\big)(f_+ - f_-)}{I_0\,(F_+ - F_-)} \\ & -\frac{1}{2}\frac{f'_+ - f'_-}{F_+ - F_-} + \frac{(f_+ - f_-)^2}{(F_+ - F_-)^2} \\ G_+ = & -\frac{1}{2}\frac{(s^+ - m_Z)f'_+ + f_+}{I_+} - \frac{(s_+ - m_Z)f_+^2}{I_+(1-F_+)} + \frac{1}{2}\frac{f'_+}{1-F_+} + \frac{f_+^2}{(1-F_+)^2} \end{cases}$$

On peut alors écrire de manière générale :

$$\forall s \in \{+, 0, -\},\ E_\epsilon[\epsilon_i | Z = z, s_i = s] = C_s\left\{1 + D_s(z - m_Z) + G_s(z - m_Z)^2\right\} + o\left((z - m_Z)^2\right)$$

On cherche alors à calculer :

$$E[\epsilon_i | \{s_j\}_{j\in\mathcal{E}}] = \int_{-\infty}^{+\infty} E_\epsilon[\epsilon_i | Z = z, s_i]\ f_Z(z | \{s_j\}_{j\in\mathcal{E}}) dz$$

On peut alors réutiliser les résultats intermédiaires de la partie 3.2 :

$$f_Z(z | \{s_j\}_{j\in\mathcal{E}}) = \frac{P\left(\{S_i\} = \{s_i\} | Z = z\right) \cdot f_Z(z)}{P\left(\{S_i\} = \{s_i\}\right)}$$

$$\begin{aligned} P\left(\{S_i\}_{i\in\mathcal{E}} = \{s_i\}_{i\in\mathcal{E}} | Z = z\right) &= (1 - F_+)^{N^+} \times (F_+ - F_-)^{N^0} \times F_-^{N^-} \\ &\times \left[1 + N\tilde{\Gamma}(z - m_Z) + (N\tilde{\Phi} + N^2\tilde{\Omega})(z - m_Z)^2\right] \\ &+ o((z - m_Z)^2) \end{aligned}$$

Avec :

$$\begin{cases} \tilde{\Gamma} &= \frac{f_+}{1-F_+}\tilde{R}^+ - \frac{f_+-f_-}{F_+-F_-}\tilde{R}^0 - \frac{f_-}{F_-}\tilde{R}^- \\ \tilde{\Phi} &= \frac{1}{2}\tilde{R}^-\left(\frac{f'_-}{F_-} - \frac{f_-^2}{F_-^2}\right) + \frac{1}{2}\tilde{R}^0\left(\frac{f'_+-f'_-}{F_+-F_-} - \frac{(f_+-f_-)^2}{(F_+-F_-)^2}\right) + \frac{1}{2}\tilde{R}^+\left(-\frac{f'_+}{1-F_+} - \frac{f_+^2}{(1-F_+)^2}\right) \\ \tilde{\Omega} &= \tilde{R}^-\tilde{R}^0\frac{f_-}{F_-}\frac{f_+-f_-}{F_+-F_-} - \tilde{R}^0\tilde{R}^+\frac{f_+-f_-}{F_+-F_-}\frac{f_+}{1-F_+} - \tilde{R}^+\tilde{R}^-\frac{f_+}{1-F_+}\frac{f_-}{F_-} \\ & \quad +\frac{1}{2}\left(\tilde{R}^-\right)^2\frac{f_-^2}{F_-^2} + \frac{1}{2}\left(\tilde{R}^0\right)^2\frac{(f_+-f_-)^2}{(F_+-F_-)^2} + \frac{1}{2}\left(\tilde{R}^+\right)^2\frac{f_+^2}{(1-F_+)^2} \end{cases}$$

$$E[\epsilon_i|\{s_j\}_{j\in\mathcal{E}}, s_i = s] = \frac{C_s\left[1+(G_s+D_sN\tilde{\Gamma}+N\tilde{\Phi}+N^2\tilde{\Omega})\sigma_Z^2\right]+o\left(\sigma_Z^2\right)}{1+(N\tilde{\Phi}+N^2\tilde{\Omega})\sigma_Z^2+o\left(\sigma_Z^2\right)}$$

$$E[\epsilon_i|\{s_j\}_{j\in\mathcal{E}}, s_i = s] = C_s\left[1+(G_s+D_sN\tilde{\Gamma})\sigma_Z^2\right] + o\left(\sigma_Z^2\right)$$

Finalement, en utilisant (12) et l'expression (6), on obtient :

$$E[T_i|\{s_j\}_{j\in\mathcal{E}}, s_i = s] = m_T + C_s(1+G_s\sigma_Z^2) + N\tilde{\Gamma}\sigma_Z^2 + C_sD_sN\tilde{\Gamma}\sigma_Z^2 + o\left(\sigma_Z^2\right)$$

On utilise $T^\epsilon = \Sigma_{i\in\mathcal{E}}\frac{q_i}{Q}T_i$ avec $\Sigma_{i\in\mathcal{E}}\frac{q_i}{Q} = 1$. L'agrégation donne alors :

$$\begin{aligned} E[T^{\mathcal{E}}|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] = \; & m_T \\ & + \textstyle\sum_{s\in\{+,0,-\}}\hat{R}^sC_s(1+G_s\sigma_Z^2) \\ & + N\tilde{\Gamma}\sigma_Z^2 \\ & + \left(\textstyle\sum_{s\in\{+,0,-\}}\hat{R}^sC_sD_s\right)N\tilde{\Gamma}\sigma_Z^2 \\ & + o\left(\sigma_Z^2\right) \end{aligned} \tag{27}$$

Pour revenir aux notations de la partie 3.1, on calcule A_+, A_0 et A_- avec la modélisation de cette partie ($T_i = Z + \epsilon_i$).

$$\forall s \in \{+,0,-\},\ A_s = E\left[Z+\epsilon_i|s_i = s\right]$$

$$\forall s \in \{+,0,-\},\ A_s = E_Z\left[Z + C_s(\left\{1 + D_s(Z-m_Z) + G_s(Z-m_Z)^2\right\} + o\left((Z-m_Z)^2\right)\right| s_i = s]$$

Si on néglige l'information apportée par la réponse s_i d'une unique entreprise i à l'estimation de la tendance commune Z, on peut alors écrire :

$$\forall s \in \{+,0,-\},\ A_s = E_Z\left[Z + C_s(\left\{1 + D_s(Z-m_Z) + G_s(Z-m_Z)^2\right\} + o\left((Z-m_Z)^2\right)\right]$$

On obtient alors :

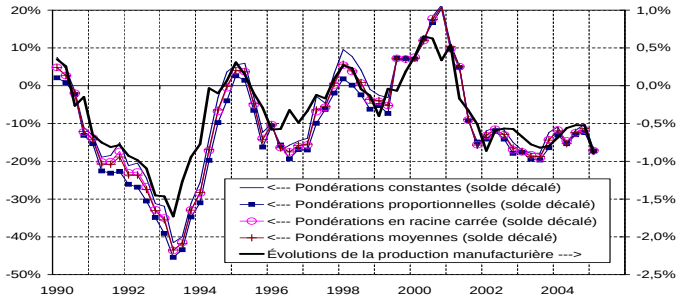
$$\forall s \in \{+,0,-\},\ A_s = m_T + C_s\left(1 + G_s\sigma_Z^2\right) + o\left(\sigma_Z^2\right)$$

Il est alors possible d'écrire :

$$\begin{aligned} E[T^{\mathcal{E}}|\mathcal{I}_{-1}, \{s_j\}_{j\in\mathcal{E}}] = \quad & m_T \\ & +\sigma_T \textstyle\sum_{s\in\{+,0,-\}} a_s \hat{R}^s \\ & +N\tilde{\Gamma}\sigma_Z^2 \\ & +\left(\textstyle\sum_{s\in\{+,0,-\}} \hat{R}^s C_s D_s\right) N\tilde{\Gamma}\sigma_Z^2 \\ & +o\left(\sigma_Z^2\right) \end{aligned}$$

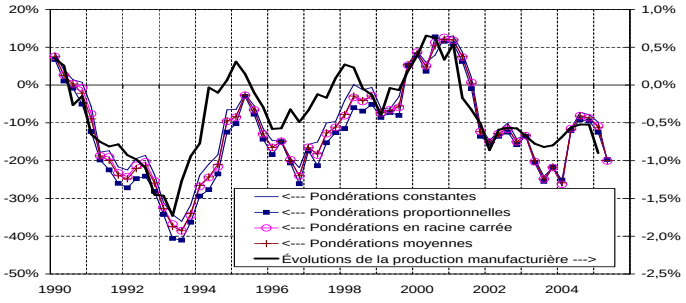
C Graphiques

FIG. 5 – Évolutions des effectifs salariés de l'industrie manufacturière et soldes relatifs aux évolutions passées des effectifs



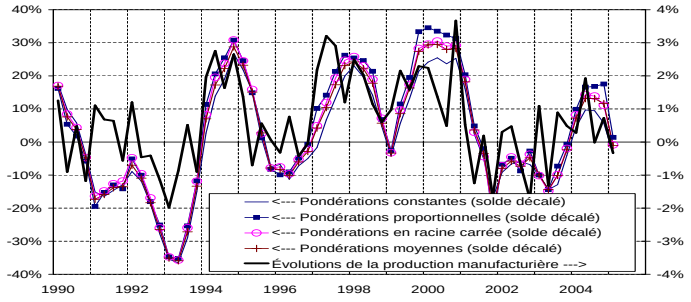
Sources : Enquête Activité et comptes trimestriels, Insee. Calculs des auteurs.

FIG. 6 – Évolutions des effectifs salariés de l'industrie manufacturière et soldes relatifs aux évolutions prévues des effectifs



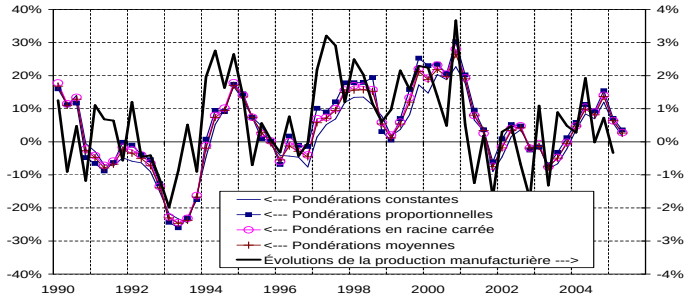
Sources : Enquête Activité et comptes trimestriels, Insee. Calculs des auteurs.

FIG. 7 – Évolutions de la production manufacturière et soldes relatifs aux évolutions passées de la production



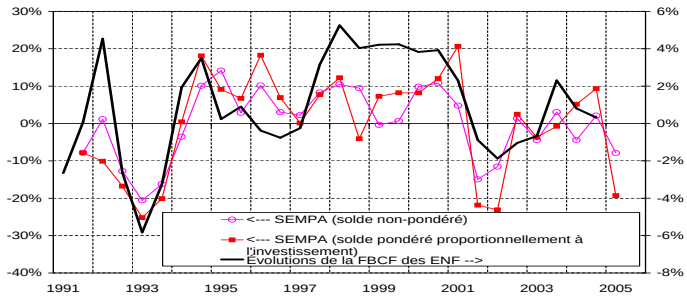
Sources : Enquête Activité et comptes trimestriels, Insee. Calculs des auteurs.

FIG. 8 – Évolutions de la production manufacturière et soldes relatifs aux évolutions prévues de la production



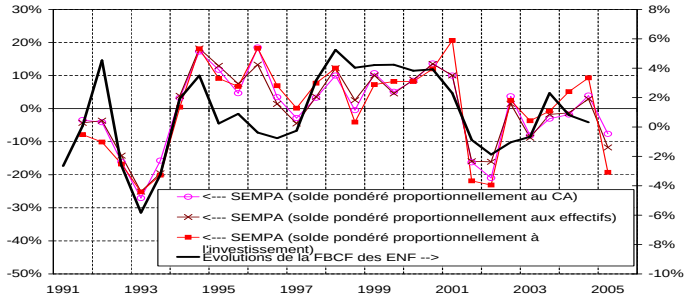
Sources : Enquête Activité et comptes trimestriels, Insee. Calculs des auteurs.

FIG. 9 – Evolutions de l'investissement et solde équipondéré versus solde pondéré par l'investissement



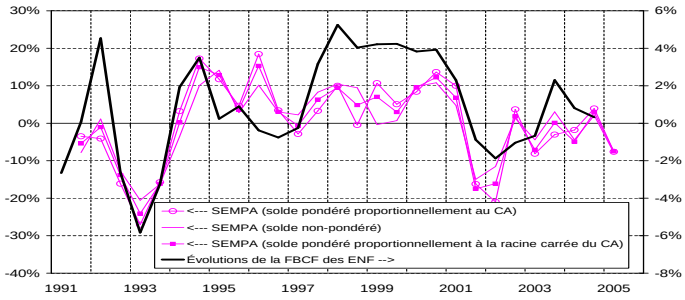
Sources : Enquête Investissement et comptes trimestriels, Insee. Calculs des auteurs.

FIG. 10 – Évolutions de l'investissement et soldes pondérés proportionnellement aux chiffres d'affaires, investissements et effectifs



Sources : Enquête Investissement et comptes trimestriels, Insee. Calculs des auteurs.

FIG. 11 – Évolutions de l'investissement et soldes pondérés de différentes manières à l'aide des chiffres d'affaires



Sources : Enquête Investissement et comptes trimestriels, Insee. Calculs des auteurs.