Tarification de pointe aéroportuaire

Marianne Raffarin

▶ **To cite this version:**

Marianne Raffarin. Tarification de pointe aéroportuaire. 2003. hal-01021524

HAL Id: hal-01021524
https://enac.hal.science/hal-01021524

Submitted on 17 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tarification de pointe aéroportuaire

Marianne RAFFARIN
Laboratoire d'Economie et d'Econométrie de l'Aérien

Mai-novembre 2003

Introduction

Depuis quelques années, l'image du transport aérien en Europe est ternie par l'ampleur des retards. Actuellement, en France, un vol sur quatre est retardé de plus de quinze minutes. Aucun voyageur aérien n'a l'assurance de la ponctualité de son vol, chacun anticipant une arrivée plus ou moins tardive par rapport à l'heure prévue. Beaucoup de passagers ne comprennent pas cette récurrence dans les retards.

Une des raisons avancées pour expliquer ces retards est liée aux aéroports. Au 1er semestre 2003, 16,3 % des retards de plus de quinze minutes ont eu pour cause les gestionnaires d'aéroport et les services de sûreté.

Mais les limites posées par les capacités aéroportuaires ne relèvent pas seulement des retards. Elles s'expriment essentiellement à travers l'attribution de créneaux pour atterrir et décoller afin de pallier à l'excès de demande par rapport à l'offre. La demande est donc rationnée en amont : les liaisons impliquant un aéroport coordonné sont soumises à l'allocation de droits permettant d'opérer à partir de cet aéroport.

Ces problèmes de capacité sont généralement traités par les économistes avec une tarification appropriée. La congestion aéroportuaire étant temporelle, du fait des fluctuations de la demande dans le temps, et l'accès aux infrastructures ne présentant pas la caractéristique de pouvoir être stocké, une tarification de pointe pourrait convenir.

Cette étude s'inscrit dans le cadre de travaux de recherche de la Direction de la Navigation Aérienne, de la Direction du Transport Aérien et de l'École Nationale de l'Aviation Civile. Elle vise à apporter à la DGAC des éléments d'appréciation sur la pertinence et la portée d'une tarification d'usage des infrastructures aéroportuaires à la pointe.

En matière de redevances d'atterrissage, le conseil de l'OACI formule cette première recommandation [1] :

> « Les redevances d'atterrissage devraient être fondées sur le poids en prenant comme base de calcul le poids maximal admissible au décollage [...] Il faudrait cependant permettre l'utilisation d'une redevance fixe par aéronef ou une combinaison d'une redevance fixe et d'un élément lié au poids, dans certaines circonstances comme aux aéroports encombrés et pendant les périodes de pointe. »

Cette recommandation de l'OACI et les exemples à travers le monde de redevances aéroportuaires fonction du temps amènent effectivement à s'interroger sur la pertinence d'une tarification de pointe pour les aéroports.

1. *Politique de l'OACI sur les redevances d'aéroport et de services de navigation aérienne*. 6e édition, 2001.

La tarification de pointe est une solution, par exemple lorsqu'il existe une demande trop forte par rapport à la capacité d'une infrastructure. Mais, elle peut aussi être utile quand il n'existe pas de contrainte sur la capacité : étant donné que les coûts fixes d'installation d'une infrastructure nécessitent une élévation du prix au-dessus du coût marginal, la tarification de pointe peut remplir efficacement ce rôle. Cependant, il est vrai que si la capacité est contrainte, le problème est plus urgent et paraît par conséquent plus évident.

La démarche dans le cadre de ce travail de recherche fut progressive, chaque étape se traduisant par un chapitre. Les deux premiers chapitres sont théoriques, les deux suivants s'intéressent plus à la pratique.

L'approche théorique de cette étude se fait par rapport à une optimisation du bien-être collectif. Un système de prix répond à des objectifs. L'objectif principal peut par exemple être de maximiser les recettes commerciales du producteur ou de maximiser le surplus social. Nous nous plaçons du point de vue d'une autorité publique, gestionnaire d'un aéroport, dont l'intérêt est d'atteindre un niveau de surplus social le plus élevé possible. Ce surplus se compose du bien-être des passagers, des profits des compagnies et du profit de l'opérateur chargé de l'aéroport.

Le premier chapitre présente la structure des prix. Nous examinons quels sont les fondamentaux des prix ; nous verrons par la suite de quelle manière les prix de pointe peuvent être intégrés dans cette structure. Les prix jouent un rôle d'optimisation. Les niveaux auxquels ils sont fixés doivent permettre de :

- bénéficier d'une infrastructure de taille optimale : c'est l'efficacité dynamique ; elle est liée à l'investissement ;
- assurer l'utilisation optimale de l'infrastructure existante : c'est l'efficacité allocative ; elle est liée à l'usage. Pour que les prix jouent ce rôle d'allocation et obtenir la meilleure qu'il soit, il faut connaître les agents qui valorisent le plus l'infrastructure ;
- encourager à une minimisation des coûts : c'est l'efficacité productive.

Il existe des *a priori* négatifs à l'encontre de la tarification de pointe : elle est souvent associée à la discrimination tarifaire. Il est vrai que la tarification de pointe peut être discriminante, mais cela ne signifie pas qu'elle est injuste ; elle peut être bonne sur le plan social. Nous cherchons à travers cette étude à lever des ambiguïtés encore fortes sur cette question.

Le second chapitre aborde la théorie de la tarification de pointe. Une telle tarification appliquée aux aéroports risque de poser problème vis-à-vis des usagers directs, que sont les compagnies aériennes. Sur un plan juridique, il existe de nombreux recours déposés par les compagnies devant les juridictions mettant en cause la légalité des évolutions des redevances. Les enjeux des conditions de fixation des redevances aéroportuaires sont juridiques, financiers et économiques. L'objet de cette étude est donc de fournir une base solide à la justification économique d'une tarification de pointe.

Il existe des modèles de prix de pointe développés par de nombreux économistes. Nous allons étudier leurs différents résultats.

La théorie distingue deux cas dans le raisonnement à adopter pour pratiquer une tarification de pointe. Elle répond soit à une logique de couverture des coûts, lorsque la demande

peut être parfaitement anticipée, soit à une logique de révélation des préférences, lorsque la demande est incertaine.

Le troisième chapitre donne des exemples de tarification de pointe dans des industries de réseau. L'expérience dans des secteurs autres que le domaine aéroportuaire peut apporter des enseignements. Les expériences d'autres aéroports sont aussi instructives. Elles illustrent la manière dont certaines difficultés liées, au passage de la théorie à la pratique, ont été levées.

Peu d'aéroports pratiquent la tarification de pointe. C'est le cas des aéroports de Londres, de New-York, de Manchester, de Toronto, et de quelques autres encore. La principale raison expliquant le recours à des redevances d'atterrissage et de décollage de pointe est la congestion ; les autorités aéroportuaires souhaitent ainsi lisser la demande et faire payer les externalités imposées.

Le quatrième chapitre traite de la tarification actuelle des deux principaux aéroports parisiens et envisage la mise en place d'une tarification de pointe. Les redevances aéronautiques, telles qu'elles sont définies pour le moment pose des problèmes sur le plan de la congestion. L'étude d'une éventuelle tarification de pointe appliquée aux aéroports parisiens nécessite de disposer de données sur ces aéroports et d'informations sur les compagnies et les passagers qui les utilisent.

Nous conclurons en donnant un résumé de tout ce qui a été présenté dans cette étude.

Chapitre 1

Accès à une infrastructure : la structure des prix

Introduction

Produire un bien ou un service engage des dépenses : c'est le coût de production ; cela se fait en échange d'une rétribution : c'est le prix. L'un des premiers rôles du prix est de récupérer les sommes engagées dans la production. Les prix sont donc en partie déterminés au regard des coûts. Ils sont aussi le reflet du rapport de force qui existe entre l'offre et la demande. Si le producteur est confronté à une concurrence rude, et qu'il ne parvient pas à vendre ses stocks, il sera amené à baisser ses prix. En revanche, en situation de monopole et face à une forte demande, il pourra tirer avantage de cette situation et pratiquer des prix élevés. Les prix servent donc avant tout à couvrir les coûts et à équilibrer l'offre et la demande.

Qu'en est-il pour les prix d'accès à une infrastructure ? Les principes sont les mêmes, mais des questions se posent de manière accrue dans ce contexte. Une infrastructure correspond à l'ensemble d'une installation indispensable pour assurer des services. Ces services visent généralement à faire circuler de l'énergie, de l'information, des biens matériels ou des personnes. C'est le cas par exemple des ports, des aéroports, des routes terrestres, des réseaux câblés pour la téléphonie fixe, pour l'électricité ou encore pour la télévision.

Ces infrastructures, mises en place pour assurer des activités de transport, sont donc des monopoles, locaux ou nationaux, et sont pour la plupart gérés par une autorité publique. Si la gestion est publique, il est important d'exercer un contrôle sur les sommes engagées car le risque que le niveau de dépenses ne soit pas efficace existe. Le caractère de monopole de ces infrastructures signifie qu'il est difficile, voire impossible, de leur trouver des substituts. Il est donc nécessaire de veiller aux coûts afin que les prix ne soient pas fixés à des niveaux trop hauts.

Ainsi, la maîtrise des coûts et le niveau de la demande jouent un rôle crucial dans la

détermination des prix d'accès à une infrastructure. Ce premier chapitre présente les déterminants fondamentaux des prix et leurs interactions.

La section 1.1 s'intéresse à la régulation économique par les prix. Les relations entre les prix et les conditions de l'offre, c'est-à-dire essentiellement les coûts, peuvent être abordées de deux points de vue : soit on considère que le rôle des prix est de couvrir les coûts, soit au contraire on assigne aux prix l'objectif d'inciter à une maîtrise des coûts.

La section 1.2 traite de la discrimination tarifaire. Les relations entre les prix et les conditions de la demande, conduisent également à envisager deux conceptions. Il est possible d'instaurer, d'une part, des subventions croisées entre les utilisateurs de l'infrastructure, en répartissant la charge financière différemment entre les utilisateurs, afin soit d'augmenter le bien-être social, soit d'augmenter le profit, et, d'autre part, un système qui vise à agir sur la demande en instaurant un prix ayant pour objectif la répartition efficace des ressources. Ces deux conceptions reviennent à distinguer la discrimination tarifaire liée à un signal exogène des agents et celle liée au choix des agents.

1.1 La régulation économique par les prix

La régulation économique signifie le contrôle exercé par une autorité sur un ou plusieurs opérateurs ayant une activité spécifique. Une autorité de régulation dispose d'un large choix de moyens pour veiller au financement d'un service qu'elle assure directement ou par l'intermédiaire d'un opérateur. Dans toute la suite, nous nous intéressons seulement au cas où l'État ne peut opérer des transferts monétaires pour financer l'activité du gestionnaire de l'infrastructure. Par conséquent, celui-ci se finance uniquement à partir des recettes engendrées par son activité.

Les théories relatives à la régulation ont beaucoup évolué depuis les années 70. La « nouvelle économie publique » s'est appuyée sur la théorie des incitations et des contrats pour développer de nouveaux modes de régulation s'appliquant notamment au cas des monopoles publics.

Nous présenterons deux types de régulation opposés : la régulation « *cost-of-service* » et la régulation « *price-cap* », avant de voir comment cette évolution s'est faite et quelles sont les recommandations de la théorie.

1.1.1 La régulation « *cost-of-service* »

Jusqu'au début des années 80, la méthode du « coût du service » a dominé. Le principe de cette méthode repose sur le fait que les revenus de l'entreprise couvrent entièrement et exactement ses coûts. Ce type de régulation autorise l'entreprise à déterminer ses prix de façon à égaliser ses recettes à ses coûts. Nous ne discutons pas particulièrement ici des situations où les infrastructures offrent plusieurs services, ce qui pose alors le problème de la ventilation des coûts communs à la fourniture de ces services.

La tarification d'une infrastructure au coût marginal est difficilement envisageable. Bien qu'elle permette d'atteindre l'optimum de premier rang, elle laisse un déficit budgétaire du

fait des coûts fixes importants d'installation de l'infrastructure. Si cette méthode est adoptée, elle nécessite donc d'opérer un prélèvement fiscal qui peut lui-même être à l'origine d'inefficacités. Il arrive cependant que des infrastructures qui ne parviennent pas à couvrir leurs coûts à partir de leurs recettes, perçoivent des aides de l'État.

La méthode de régulation au coût du service la plus fréquemment citée, est celle qui égalise les prix aux coûts moyens. Le coût d'installation de l'infrastructure est ainsi réparti uniformément sur l'ensemble des usagers. Pourtant, elle ne s'est quasiment jamais appliquée du fait de son caractère d'inefficacité. La théorie économique a contribué à apporter d'autres solutions visant à assurer la couverture des coûts et à être plus efficace.

1.1.1.1 Tarification RAMSEY-BOITEUX

Un optimum de second rang peut être obtenu en ajoutant au programme de maximisation du bien-être social la contrainte budgétaire. La résolution de ce programme aboutit à une tarification RAMSEY-BOITEUX [2]. En considérant soit que l'opérateur de l'infrastructure fait payer la fourniture d'au moins deux services différents, soit qu'il est en mesure d'identifier au moins deux types d'utilisateurs différents, cette règle de prix indique que le prix doit s'élever au dessus du coût marginal de façon inversement proportionnelle à l'élasticité-prix de la demande. Le coefficient de proportionnalité est déterminé de manière à ce que la contrainte budgétaire soit satisfaite. La tarification RAMSEY-BOITEUX s'applique, par exemple, pour les services de la SNCF offerts aux passagers. Il existe des subventions croisées entre les liaisons : pour les liaisons très fréquentées les prix sont supérieurs au coût marginal ; cette situation permet de couvrir le déficit généré par des prix fixés au-dessous du coût marginal pour des liaisons peu fréquentées.

En règle générale, l'augmentation des prix fait diminuer la demande, ce qui est mauvais pour l'objectif de maximisation du surplus social. Or, face à l'obligation d'augmenter les prix pour satisfaire la contrainte budgétaire, l'idée est de majorer les prix de la demande dont l'élasticité est la plus faible, afin de ne pas trop faire chuter la demande. Ainsi, pour une élasticité-prix de la demande faible, l'inverse de cette élasticité sera forte et le prix sera bien supérieur au coût marginal. À l'opposé, une élasticité de la demande élevée aura pour conséquence un prix proche du coût marginal. Cette tarification revient à pratiquer des subventions croisées entre la production de plusieurs biens ou entre plusieurs utilisateurs. Cependant, il existe un risque que cette tarification aille à l'encontre d'une certaine équité entre les utilisateurs, puisque les biens pour lesquels l'élasticité-prix de la demande est faible sont souvent des biens indispensables (biens dits de première nécessité).

1.1.1.2 Tarif binôme

Une autre solution, qui permet d'atteindre l'optimum de premier rang est de fixer le prix égal au coût marginal et d'instaurer le paiement d'un abonnement répartissant la charge des coûts fixes sur l'ensemble des utilisateurs. Le prix de l'accès à l'infrastructure incluant l'abonnement risque d'être trop élevé pour certains utilisateurs dont les capacités financières sont limitées.

Le régulateur peut donc aussi assigner à l'opérateur de l'infrastructure un objectif d'équité au tarif binôme. Cette obligation se traduit alors par un abonnement moins élevé et un prix d'autant moins supérieur au coût marginal que l'élasticité-prix de la demande est forte. En diminuant l'abonnement au détriment du prix unitaire, il n'est plus dissuasif de demander à utiliser l'infrastructure. Cette situation était celle de France Télécom avant l'ouverture à la concurrence. Afin que le téléphone fixe soit accessible au plus grand nombre, l'opérateur offrait un abonnement à un montant inférieur à son niveau de premier rang et fixait un prix unitaire supérieur au coût marginal des télécommunications. Ce mécanisme, intéressant pour ces effets redistributifs, n'est cependant pas optimal.

1.1.1.3 Inconvénients de la régulation « *cost-of-service* »

La régulation « *cost-of-service* » a l'avantage de garantir l'existence de certains services publics en leur évitant la faillite. Mais, elle pose le problème de l'absence d'incitations. L'opérateur étant assuré de récupérer tous ses frais engagés, il n'est pas incité à maîtriser ses coûts. De plus, ce mode de régulation ne prend pas en considération les différents niveaux de qualité puisque, quel que soit son choix, il n'a pas d'incidence sur son profit. L'opérateur n'est donc pas incité à atteindre le niveau de qualité optimal ; il peut choisir une qualité trop élevée ou trop faible.

1.1.2 La régulation « *price-cap* »

Les problèmes soulevés par le manque d'incitations et l'existence d'aléa moral ont conduit à développer une nouvelle méthode de régulation, qualifiée de « *price-cap* ». C'est un mécanisme de prix plafonnés : le régulateur fixe le taux maximum auquel l'opérateur peut accroître ses prix. Ce taux est généralement la différence entre le taux d'inflation (RPI, *Retail Price Index*) et le taux de croissance de la productivité dans le secteur considéré (X), fixé contractuellement entre le régulateur et l'opérateur. Cette règle, couramment appelée « *RPI-X* », établit donc un plafond d'évolution des prix.

1.1.2.1 Incitation à la maîtrise des coûts

Les revenus de l'opérateur sont donc totalement indépendants des coûts engagés. C'est cette indépendance qui incite fortement l'opérateur à maîtriser ses coûts. Il peut dégager des profits d'autant plus importants qu'il parvient à creuser l'écart entre ses recettes et ses coûts réels. L'opérateur a donc intérêt à avoir des coûts inférieurs à la somme qu'il récupérera, grâce à une productivité supérieure à celle fixée contractuellement.

Avec la régulation précédente, on avait $R_{cs}(C_{cs}) = C_{cs}$, où C_{cs} sont les coûts de l'opérateur et $R_{cs}(C_{cs})$ ses revenus établis de façon à couvrir les coûts. Ceci signifiait que le profit $(\Pi(C_{cs}))$ était nul, quel que soit le niveau des coûts. À l'inverse, la régulation « *price-cap* » est à l'origine de revenus fixes, indépendants des coûts : $R_{pc}(C_{pc}) = \overline{R_{pc}}$. Ainsi la maximisation du profit équivaut à la minimisation des coûts :

$$\text{Max } \Pi(C_{pc}) = \overline{R_{pc}} - C_{pc} \Longleftrightarrow \text{Min } C_{pc}$$

1.1.2.2 Problèmes soulevés par la régulation « *price-cap* »

Outre la détermination du niveau de prix initial dans le cadre d'une régulation « *price-cap* », les principaux problèmes que soulève ce mécanisme sont liés à la qualité des services, aux rentes et à la répétition des contrats. La réduction des coûts par l'opérateur risque de se faire au détriment de la qualité. Une solution peut être de compléter le contrat en fixant également un niveau de qualité.

La réduction des coûts est obtenue, avec ce système, en contrepartie d'une rente informationnelle laissée à l'opérateur. Il est important de noter à quel point la question du niveau auquel la variable X est fixée est cruciale. C'est en partie en fonction du choix de X que l'opérateur va être en mesure de dégager une forte rente ou au contraire va être face à une obligation trop forte pour pouvoir la satisfaire. Étant donné que laisser une rente peut coûter cher, il existe donc un arbitrage à faire entre les incitations à l'efficacité et l'abandon de rentes.

Dans une perspective de long terme, où les contrats fixant le niveau des gains anticipés de productivité sont renouvelés, le caractère incitatif de ce mode de régulation est moindre. Le risque pour l'opérateur, s'il révèle beaucoup d'information à l'occasion du premier contrat, est que le régulateur s'en serve pour réduire sa rente avec les contrats suivants. L'opérateur anticipant ce comportement du régulateur, peut alors modifier le sien dès le premier contrat en freinant sa révélation d'information. Pour éviter cet effet de cliquet, l'opérateur doit être persuadé que le régulateur n'abusera pas de son pouvoir en resserrant le plafonnement au fur et à mesure que l'opérateur réalise des gains de productivité.

1.1.2.3 Expérience britannique

Le recours à la régulation « *price-cap* » est fréquent pour les agences de réglementation britanniques. Ainsi, de nombreuses infrastructures voient leur prix d'accès indexés annuellement sur le taux d'inflation duquel est soustrait un facteur reflétant les gains de productivité attendus.

À l'usage, cette formule est cependant apparue insuffisante. Un autre paramètre est souvent ajouté. Il représente des facteurs spécifiques comme l'évolution des coûts d'approvisionnement ou les dépenses d'investissement pour la qualité.

C'est le cas, par exemple, des services du contrôle aérien britannique (NATS, *National Air Traffic Service*) soumis à l'autorité de la CAA (*Civil Aviation Authority*). Pour les coûts liés au service fourni aux compagnies, la règle est celle du plafond de prix, complétée par une incitation à fournir un service de bonne qualité, la maîtrise des retards. Le taux unitaire soumis au *price-cap* ne peut augmenter plus qu'une quantité M se définissant de la manière suivante :

$$M = SC + K - S$$

La partie SC correspond à la partie usuelle de l'incitation à la réduction des coûts, mesurant l'écart entre l'inflation et le taux de croissance de la productivité fixé à 3 % pour 2002, 4 % pour 2003 et 5 % pour 2004 et 2005.

Le taux unitaire est déterminé en fonction de l'anticipation du volume de trafic exprimé en termes d'unités de service pour l'année à venir. Le terme K sert à corriger l'écart dû à une

erreur de prévision, par rapport au volume de trafic effectif.

Le dernier terme, S, est relatif à la qualité. Il indique la contre-performance de l'opérateur par rapport à l'objectif du régulateur exprimé en termes de retards. Si l'opérateur a fait mieux que ce que souhaitait le régulateur, S est négatif, permettant alors à l'opérateur d'augmenter plus ses prix. À l'inverse, si les retards sont plus élevés que ce que le régulateur était prêt à accepter, l'opérateur verra son taux unitaire autorisé réduit.

1.1.3 Le principe de la régulation par les prix

Les monopoles publics ont toujours été soumis à la régulation d'une autorité publique. Cette autorité, longtemps exercée par le gouvernement via le ministre de tutelle pour l'activité considérée, est de plus en plus confiée à une entité indépendante.

Cette réorganisation des autorités publiques, au travers d'une séparation entre le gouvernement et le monopole historique, et une autorité de réglementation spécialisée s'expliquent par plusieurs raisons. La principale d'entre elles est liée aux interventions de l'État dans la gestion du service. Sans séparation, le risque est que l'État utilise le monopole à des fins politiques en lui assignant des rôles de redistribution sociale, de levier pour la relance économique et l'emploi, voire à des fins électorales, limitant alors l'efficacité du monopole et liant directement sa stratégie de développement aux aléas de la conjoncture politique.

À l'inverse le rôle de l'autorité de régulation de monopoles naturels ou d'oligopoles, outre de veiller au respect des principes de la concurrence, est de surveiller les prix et d'inciter à la maîtrise des coûts. Afin d'assurer sereinement ce rôle, le régulateur est une entité indépendante. Le régulateur atteint ces objectifs qui lui sont fixés, en passant des contrats avec le ou les opérateurs.

1.1.3.1 Évolution du mode de régulation

Le rôle du gouvernement, avant ces récentes évolutions résidait essentiellement dans la tarification administrée. Il devait fixer le prix de vente des produits du monopole. Dans la plupart des cas, les prix étaient déterminés de manière à couvrir strictement les coûts : c'est la régulation « *cost-of-service* ». La principale critique adressée à ce type de régulation est la non maîtrise des coûts. En assurant au monopole la couverture totale de ses coûts, l'autorité n'encourage pas l'opérateur à avoir une gestion efficace de ses dépenses. Cette régulation du gouvernement n'est alors pas contraignante pour le monopole public, elle va dans le sens des intérêts propres de celui-ci.

Depuis quelques années, de nombreuses questions se posent par rapport à l'intervention publique : son domaine d'intervention, les fins auxquelles elle est mise en œuvre, les coûts, directs ou indirects, qu'elle induit. C'est en quelque sorte une prise de conscience des effets pervers auxquels l'intervention publique peut conduire, qui est à l'origine de développements récents sur des modes de régulation alternatifs à la tarification administrée. Cet arbitrage entre les effets bénéfiques et les défauts de la réglementation tient au fait qu'il existe une asymétrie d'information entre le monopole et son autorité de régulation. Cette asymétrie étant en faveur du monopole, il peut profiter de cette situation pour chercher à satisfaire ses propres intérêts

plutôt que ceux de la société dans son ensemble.

1.1.3.2 Anti-sélection et aléa moral

L'information cachée par le monopole public à l'autorité de régulation concerne généralement les coûts. Il s'agit donc d'un problème de contrôle d'un agent, qualifié dans la théorie des contrats de « *Principal* », sur un autre, l'« *Agent* ». La dissimulation de l'origine de coûts trop élevés a deux causes, qui peuvent intervenir séparément ou simultanément. La difficulté du contrôle est liée soit à la non observabilité des caractéristiques de l'Agent par le Principal, lorsque l'information est exogène à l'Agent, ce qui correspond à un problème « d'anti-sélection », soit à la non vérifiabilité des actions de l'Agent par le Principal, lorsque l'information est endogène à l'Agent, ce qui correspond à un problème « d'aléa moral », soit aux deux.

Dans le cas du monopole public, l'information d'un état caché du problème d'anti-sélection concerne les caractéristiques de production. L'entreprise peut manipuler l'information sur ses coûts en déclarant que son paramètre de productivité, inconnu du régulateur, est plus élevé qu'il ne l'est en réalité, signifiant alors un état peu productif. Le problème d'aléa moral, celui d'une décision cachée, concerne les choix d'investissement et les efforts de gestion de l'opérateur réglementé. Les actions entreprises par l'opérateur pour réduire les coûts sont inconnues du régulateur.

C'est l'existence de contraintes essentiellement informationnelles qui empêchent le régulateur d'imposer sa politique idéale, celle qui aboutirait à une situation de premier rang. La nouvelle économie publique suggère que le Principal mette en place des contrats incitatifs, qui conduisent l'Agent à adopter un comportement conforme à celui qu'il souhaite. Par exemple, un employeur peut conditionner les salaires de ses employés aux résultats de l'entreprise, ou bien un assureur peut instaurer un système de franchise en cas de responsabilité des assurés dans un sinistre.

On considère généralement que le régulateur est en mesure d'observer le résultat, mais qu'il n'est pas capable de juger ce qui provient de la caractéristique exogène et ce qui provient de l'action endogène. Lorsque l'asymétrie d'information porte sur le coût, le régulateur observe quelles ont été les dépenses engagées. Mais il ne sait pas démasquer l'opérateur qui fait passer un niveau d'effort sous-optimal pour une dotation technique peu performante. Le régulateur qui souhaite une réduction des coûts peut donc l'obtenir, au moins en partie, en échange de l'abandon d'une rente au profit du régulateur. Cependant, il est dans l'intérêt collectif que cette rente ne soit pas trop élevée.

À l'objectif de maximisation du bien-être social du régulateur s'ajoutent deux autres buts, ceux d'inciter à une réduction des coûts et d'extraire la rente de l'opérateur.

En présence de la double asymétrie d'information, anti-sélection et aléa moral, lorsque l'un des buts est atteint, c'est nécessairement au détriment de l'autre.

Ainsi, sans informations sur la productivité et le niveau d'effort de l'opérateur, le régulateur ajoute des contraintes à son programme de maximisation du bien-être social. Ces contraintes visent à inciter l'opérateur à révéler de l'information.

1.1.3.3 La régulation optimale

À chaque origine d'asymétrie d'information prise individuellement correspond un meilleur mode de régulation. En imaginant que l'asymétrie d'information entre le régulateur et l'opérateur soit uniquement un problème d'anti-sélection, l'opérateur dissimulant son véritable paramètre de productivité, le contrat optimal entre l'Agent et le Principal est de type « *cost-of-service* ». En l'absence d'aléa moral, il est inutile de chercher à inciter à réduire le coût et par conséquent ce contrat est le seul à ne laisser aucune rente.

En revanche, si l'asymétrie d'information entre l'opérateur et le régulateur ne portait que sur une action endogène, le meilleur contrat face à cette situation d'aléa moral serait du type « *price-cap* ». Ainsi, l'effort pour réduire les coûts serait maximal, mais il s'obtiendrait au prix de la rente la plus élevée.

Mais lorsque l'aymétrie d'information est double, les deux types de régulation ne sont pas exclusives l'un de l'autre : le mode de régulation optimal est une combinaison des deux précédents, car deux objectifs sont poursuivis. Le contrat est donc « à la mesure » de l'information.

LAFFONT et TIROLE (1993) [10] montrent que l'optimum pour le régulateur consiste à proposer à l'opérateur un menu de mécanismes de prix linéaires et de le laisser sélectionner la règle de prix qu'il préfère parmi tous les mécanismes possibles combinant des prix qui couvrent les coûts et des prix dont l'évolution est plafonnée.

En pratique, le régulateur décentralise peu le mécanisme de prix. C'est plutôt lui qui impose la manière dont l'opérateur va pouvoir déterminer ses prix. Cependant, ce choix du régulateur doit prendre en compte le degré et l'origine de l'asymétrie d'information, afin d'avoir des prix en relation avec les coûts, reflétant convenablement la situation.

1.2 La discrimination tarifaire

Il est difficile de donner une définition unique de la discrimination tarifaire. *A priori*, un opérateur discrimine par les prix lorsqu'un même service est vendu, à des prix différents, à un seul ou à plusieurs utilisateurs. En réalité, il existe aussi une discrimination tarifaire lorsque des utilisateurs payent le même prix, alors que les coûts de fourniture de ces services sont différents.

Ainsi, nous pouvons considérer, comme le suggère PHLIPS (1988) [14], qu'il y a discrimination lorsque des variétés d'un même service lié à une infrastructure sont vendues à des utilisateurs différents à des prix nets différents. Les prix nets sont les prix payés par les utilisateurs après déduction de tous les coûts liés au service. Si la différence de prix n'est pas justifiée par une différence de coûts, alors on parle de discrimination tarifaire.

C'est le cas par exemple de la navigation aérienne. Pour profiter du même service de contrôle aérien au cours d'un même trajet, deux avions de masses différentes acquitteront des redevances différentes. Par ailleurs, deux avions de masses égales, parcourant une distance égale payeront le même montant de redevances qu'ils aient bénéficié du contrôle français, par exemple au dessus de la métropole (espace aérien très fréquenté) ou au dessus des Antilles

(espace aérien peu fréquenté), alors que les coûts de ces deux services sont différents, ceux du second étant plus élevés que ceux du premier.

Après avoir vu les conditions de discrimination d'un gestionnaire d'une infrastructure auprès de ses utilisateurs, nous reviendrons en détail sur la discrimination liée à un signal exogène aux agents et sur celle liée au choix des agents.

1.2.1 Le principe de la discrimination tarifaire

Afin de voir sur quel principe repose la discrimination tarifaire, il est nécessaire d'examiner, d'une part, de quelles manières il est possible de discriminer, d'autre part, quelles sont les conditions requises pour être en mesure de discriminer.

1.2.1.1 Taxinomie de Pigou

En 1920, PIGOU [15] a proposé une taxinomie de la discrimination par les prix : une classification des pratiques discriminantes en trois degrés. La discrimination au premier degré différencie les utilisateurs un à un. Celles au second et au troisième degré distinguent les utilisateurs en fonction de caractéristiques qui leur sont endogènes ou exogènes. Nous les présentons rapidement ici, et reviendrons en détail sur les deux dernières dans la suite de la section.

La discrimination par les prix au premier degré, dite « *parfaite* », signifie que l'opérateur vend un bien à un prix différent à chaque utilisateur. Pour chacun, le prix est égal à sa disponibilité à payer. Face à de tels prix, les utilisateurs sont indifférents entre ne pas utiliser l'infrastructure ou l'utiliser aux conditions de l'opérateur discriminant. Afin de s'assurer d'un choix en faveur de l'utilisation de l'infrastructure, l'opérateur doit fixer un prix légèrement inférieur à la valeur de réserve des agents. L'opérateur s'approprie de cette manière la totalité du surplus des utilisateurs de son infrastructure. L'étude des conséquences en termes de bien-être collectif de cette politique est facile à réaliser. Le programme du producteur consiste à maximiser ses profits, sous la contrainte que les agents utilisent l'infrastructure. À partir de là, il n'existe pas de moyens d'améliorer la situation d'un agent sans affecter celle d'un autre. Le profit ne peut s'accroître puisqu'il est déjà à son maximum, et le surplus des utilisateurs ne peut être augmenté sans diminuer le profit de l'entreprise. Ainsi, la discrimination au premier degré est PARETO-optimale. La discrimination parfaite est donc généralement associée au fait que tous les surplus individuels sont capturés par l'opérateur. Dans la réalité, cette politique tarifaire est très difficile à mettre en œuvre. L'incomplétude de l'information empêche les gestionnaires d'infrastructure en position de discriminer de connaître exactement les prix de réserve des utilisateurs.

La discrimination au second degré consiste à offrir les mêmes services à différents utilisateurs, en leur proposant des prix variant avec le niveau quantitatif ou qualitatif des services. Les combinaisons prix-quantités ou prix-qualité, offertes sont déterminées de façon à ce que les utilisateurs achètent exactement la quantité ou la qualité qu'ils souhaitent, et non pas un niveau inférieur. Différents niveaux de qualité sont une manière de faire révéler aux utilisateurs leurs disponibilités à payer pour l'utilisation d'une infrastructure.

Dans la discrimination au troisième degré, l'opérateur propose un même service lié à l'infrastructure à différents prix, en divisant la demande totale en plusieurs « marchés » sur la base d'une information exogène. Il s'agit de signaux directement observables et non, comme dans le cas précédent, du résultat de choix proposés aux acheteurs. Si l'opérateur est en mesure d'identifier des caractéristiques spécifiques des utilisateurs, corrélées avec leurs disponibilités à payer, cette forme de discrimination est la plus simple à mettre en œuvre.

1.2.1.2 Conditions pour discriminer

Afin de pratiquer une discrimination tarifaire, trois conditions doivent être réunies : avoir un pouvoir de marché, pouvoir trier les utilisateurs et éviter l'arbitrage.

Premièrement, la discrimination est une politique de prix obligatoirement liée à l'existence d'un pouvoir de marché, le cas extrême étant celui du monopole. Pour discriminer, les entreprises doivent être assurées de ne pas risquer de perdre leur clientèle. Les agents qui subissent négativement cette discrimination auraient intérêt à adresser leur demande aux concurrentes de l'entreprise discriminante. Pour l'opérateur, gestionnaire d'une infrastructure, ce risque est limité, voire inexistant. Par exemple, une compagnie aérienne qui souhaite desservir une certaine région, peut faire jouer la concurrence entre plusieurs aéroports. Ce comportement est pourtant très limité : le nombre d'aéroports dans une région est restreint et tous ne disposent pas de l'équipement nécessaire pour certains avions.

Une deuxième condition à la mise en œuvre de pratiques discriminatoires est qu'il soit « possible » de trier les utilisateurs ; ce qui renvoie à deux notions : celle « d'en être capable » et « d'y être autorisé ». Lorsque l'opérateur est face à des demandes qui diffèrent selon des types exogènes, il lui est facile de discriminer. En revanche, lorsque les origines diverses des demandes ne s'expliquent plus à l'aide de telles caractéristiques, il est souvent plus difficile d'identifier « différents marchés ». L'opérateur discrimine alors en fonction de types endogènes, résultats d'un choix de la part des utilisateurs. Il fait alors dépendre le prix, de caractéristiques du service qui permettent de faire révéler aux utilisateurs leur disponibilité à payer pour l'utilisation de l'infrastructure. Il peut s'agir, par exemple, de la période à laquelle l'achat est fait ou d'attributs qui, apportés au bien, l'améliore. Pour pratiquer différents prix selon les utilisateurs, il faut également que les dispositions légales l'autorisent. Par exemple, les prix de France Télécom pour que d'autres opérateurs de téléphonie fixe utilisent l'infrastructure sont soumis aux règles de l'Autorité de Régulation des Télécoms.

Une troisième condition à la mise en œuvre de pratiques discriminatoires par les prix est de pouvoir empêcher l'exercice d'un arbitrage. Un agent est en mesure d'arbitrer s'il peut obtenir le bien qu'il désire à deux prix différents. L'arbitrage est alors le résultat d'une transférabilité qui peut concerner les biens ou les demandes. La transférabilité de biens signifie la possibilité pour certains agents de les acheter non pas directement au producteur, mais par l'intermédiaire d'autres agents. Il existe des biens non transférables par nature. Associés à une entité ou à un lieu, des services fournis dans le cadre d'une infrastructure ne peuvent être vendus à une entité et utilisés par une autre. Le second type de transférabilité est celui de la demande, émanant d'un même agent, entre différents biens. L'opérateur doit s'assurer que chaque utilisateur acquiert bien ce qui lui est destiné. Cette condition rejoint la capacité

de l'opérateur à trier les utilisateurs. Lorsque les prix reposent sur des caractéristiques identifiables des utilisateurs, la demande n'est absolument pas transférable. Le problème de la transférabilité se pose lorsque le vendeur ne fait plus face à des particularités exogènes de l'utilisateur mais endogènes. L'opérateur doit s'assurer que la différence de prix ne fait pas plus que compenser la réduction de la qualité ou la quantité.

Nous étudions maintenant les manières de discriminer au second et au troisième degré, ainsi que leurs conséquences en termes d'optimalité et de surplus.

1.2.2 La discrimination tarifaire liée à un signal exogène aux agents

La discrimination tarifaire liée à un signal exogène des agents repose sur la séparation de la demande totale en plusieurs « marchés » sur la base d'une information exogène, par l'opérateur de l'infrastructure. Il s'agit de signaux directement observables et non du résultat de choix proposés aux acheteurs.

Cette forme de discrimination correspond à une segmentation du marché ; c'est la plus facile à mettre en œuvre. De plus, elle peut satisfaire à la fois des intérêts privés et l'intérêt collectif. Pour le montrer nous nous intéressons d'abord à un opérateur maximisant son profit, puis à un opérateur maximisant le bien-être social.

1.2.2.1 Maximisation du profit

Afin de diviser la demande qui s'adresse à lui en plusieurs groupes, l'opérateur doit identifier des caractéristiques particulières de ses utilisateurs sur lesquelles il va faire reposer sa discrimination. Ces profils doivent être tels que des utilisateurs appartenant à un même groupe présentent une fonction de demande individuelle similaire. Ensuite, au sein de chaque groupe, il n'existe pas de discrimination. En identifiant m « marchés », l'opérateur discriminant propose sur les différents marchés les prix $\{p_1, \ldots, p_m\}$ et les quantités demandées associées à ces prix sont $\{q_1 = D_1(p_1), \ldots, q_m = D_m(p_m)\}$. Le programme de l'opérateur est donc de maximiser :

$$\Pi(p_1, \ldots, p_m) = \sum_{i=1}^{m} p_i D_i(p_i) - C\left(\sum_{i=1}^{m} D_i(p_i)\right) \tag{1.1}$$

Cette situation est équivalente à celle qui consiste à vendre plusieurs biens de nature différentes, pour lesquels il existe des coûts communs. Il s'agit d'un problème de tarification multi-produits. La maximisation du profit par rapport au vecteur de prix, variables d'ajustement dont se sert le producteur pour atteindre son objectif, conduit à la règle de l'élasticité inverse :

$$\begin{aligned} \frac{\partial \Pi}{\partial p_i} &= 0, \forall i \\ \Rightarrow D_i(p_i) + p_i \frac{\partial D_i}{\partial p_i} - \frac{\partial C}{\partial q}\frac{\partial D_i}{\partial p_i} &= 0, \forall i \end{aligned}$$

$$\Rightarrow p_i - \frac{\partial C}{\partial q} = -\frac{D_i(p_i)}{\frac{\partial D_i}{\partial p_i}}, \forall i$$
$$\Rightarrow \frac{p_i - \frac{\partial C}{\partial q}}{p_i} = \frac{1}{\epsilon_i}, \forall i$$

où ϵ_i est l'élasticité de la demande d'un utilisateur du groupe i par rapport au prix sur le marché i.

Ainsi, l'opérateur discriminant fait payer plus cher les groupes dont l'élasticité de la demande est plus faible. Les utilisateurs à faible élasticité de la demande modifient peu leur demande face à une variation de prix. L'opérateur peut donc plus facilement augmenter leurs prix. En revanche, pour ceux plus sensibles aux hausses de prix, il n'est pas dans l'intérêt de l'opérateur d'augmenter leurs prix, il risquerait de les voir fortement diminuer leur demande.

On retrouve une tarification recommandée dans le cadre d'une régulation « *cost-of-service* » (cf. section (1.1.1)) : la tarification RAMSEY-BOITEUX.

Les prix d'opérateurs privés reposant par exemple sur l'âge de leurs utilisateurs illustrent ce type de discrimination. C'est le cas de la compagnie aérienne Air France qui proposent des tarifs « jeunes ».

1.2.2.2 Optimalité et surplus

Afin de mesurer les conséquences d'une discrimination au troisième degré sur le bien-être social, nous comparons les surplus entre la situation où l'opérateur pratique un prix uniforme et celle où il propose différents prix selon des caractéristiques exogènes des utilisateurs. À quantité fixe, la discrimination tarifaire est socialement moins bonne qu'une tarification uniforme. La condition nécessaire pour que la discrimination au troisième degré domine un prix uniforme, (ROBINSON (1933) [17] et SCHMALENSEE (1981) [18]) est que la production totale augmente. Ce résultat est illustré sur la figure (1.1).

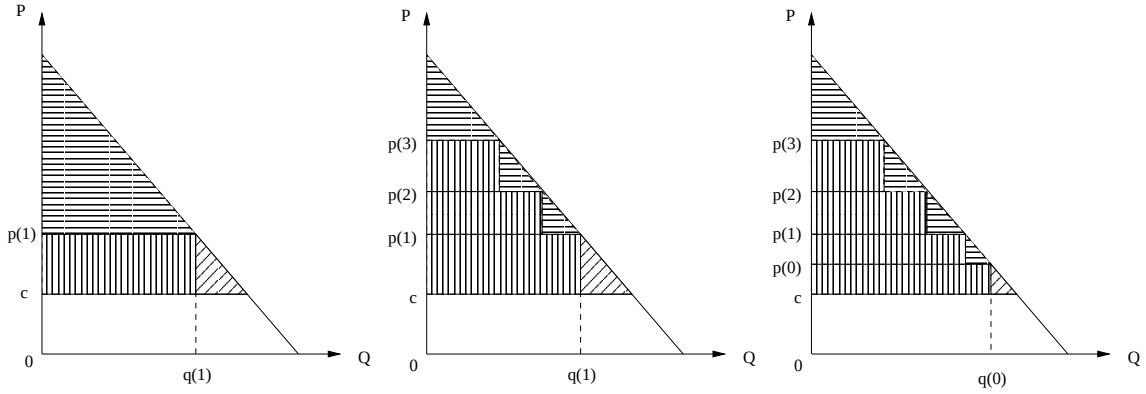

FIG. 1.1 – *Comparaison des surplus entre prix uniforme et discrimination tarifaire.*

Le premier cas correspond à la situation d'un prix uniforme, le second à celle d'une discrimination tarifaire sans augmentation de la quantité, le troisième à celle d'une discrimination

tarifaire avec une hausse de la quantité totale.

On observe que le profit de l'opérateur, partie hachurée verticalement, est plus élevé dans la seconde situation que dans les deux autres. Son profit a augmenté, d'une part, car le nombre de utilisateurs a accru et, d'autre part, car les utilisateurs disposés à payer plus cher doivent acquitter un prix plus élevé. Dans la troisième configuration, il est tout de même plus élevé que dans la première.

Pour les utilisateurs, partie hachurée horizontalement, la première situation leur est plus favorable que la seconde. Mais le bien-être social, avec un poids identique pour les utilisateurs et l'opérateur, est le même dans les deux premiers cas, car la quantité échangée est strictement identique. En revanche, dans le dernier cas, le surplus collectif augmente. la perte sèche, partie hachurée obliquement, liée au fait que certains utilisateurs n'ont pas accès à l'infrastructure, car le prix est supérieur au coût marginal c, est réduite. L'introduction d'un prix en-dessous du prix $p(1)$ permet à des utilisateurs qui n'entraient pas sur le marché de le faire. L'entrée de ces nouveaux utilisateurs augmentent le surplus social. La discrimination peut donc avoir des effets positifs sur le plan du bien-être social.

Cette discrimination, en vue de maximiser le surplus social, peut aussi être appliquée à travers une tarification RAMSEY-BOITEUX. C'est le cas des redevances aériennes payées par les compagnies pour bénéficier des services de la navigation aérienne. Ces redevances sont fonction de la masse de l'avion. La justification des autorités, nationales et internationales, pour ce système de prix repose sur une corrélation supposée entre la taille des avions et la capacité financière de la compagnie pour le vol. Le poids de l'avion est dans ce cas considéré comme une caractéristique exogène aux compagnies qui reflète leur disponibilité à payer pour le service (RAFFARIN (2002) [16]).

La discrimination tarifaire au troisième degré est mise en œuvre par des opérateurs dont les objectifs sont divergents. Les uns ont pour préoccupation le bien-être social et pratiquent la discrimination tarifaire dans un souci de redistribution, les autres cherchent à maximiser leur profit sans intention redistributive et sont également amener à discriminer.

1.2.3 La discrimination tarifaire liée au choix des agents

La discrimination tarifaire liée au choix des agents consiste pour l'opérateur, à proposer plusieurs couples prix-quantités ou prix-qualité aux utilisateurs de l'infrastructure et de les laisser choisir le couple qui leur convient le mieux. Le prix unitaire du service est donc fonction de la quantité achetée ou dépend de ses attributs par rapport à ceux d'un autre service lié à l'infrastructure.

Étant donné les possibilités d'arbitrage personnel, le programme de l'opérateur maximisant son profit comporte des contraintes d'auto-sélection, dites « *incitatives* ». Le but de ces contraintes est de faire en sorte que les utilisateurs achètent le lot ou la qualité qui leur est destiné. Le résultat n'est pas nécessairement optimal du point de vue du surplus social.

1.2.3.1 Maximisation du profit

La plus courante des discriminations tarifaires par les quantités est la tarification binôme. La règle de prix se représente dans un repère quantité (q)-dépense (T) par une droite non linéaire mais affine. Quelque soit son niveau d'utilisation de l'infrastructure, chaque agent doit payer une somme fixe A. La seconde partie de la dépense, dépend du niveau d'utilisation de l'infrastructure, chaque unité étant vendue au prix unitaire p. Ainsi, la dépense totale s'écrit : $T = A + pq$. La dépense moyenne est décroissante avec la quantité achetée, ce tarif correspond alors à une remise sur les quantités.

Il existe des mécanismes de prix non linéaires plus complexes que les tarifs binômes. L'opérateur discriminant propose aux utilisateurs des couples prix-quantités ou prix-qualité.

Soient deux catégories d'utilisateurs de type t_1 et t_2, respectivement en proportion f_1 et f_2 dans la population et dont la fonction d'utilité est $u(x_1, t_1)$ et $u(x_2, t_2)$, avec x_i le niveau d'utilisation de l'infrastructure ou le niveau de qualité du service choisi par un agent de type t_i. Nous décidons de considérer que x_i correspond à une quantité. Les hypothèses sur ces fonctions sont :

$$u(x, t_2) > u(x, t_1)$$

$$\frac{\partial u(x, t_2)}{\partial x} > \frac{\partial u(x, t_1)}{\partial x}$$

Ces inégalités signifient que les agents de type t_2 et l'utilisateur marginal de type t_2 sont plus disposés à payer pour une quantité donnée que respectivement les agents de type t_1 et l'utilisateur marginal de type t_1. Les agents de type t_2 sont des utilisateurs à demande élevée alors que les agents de type t_1 ont des demandes faibles. L'opérateur va proposer deux couples (r_1, x_1) et (r_2, x_2), où r_i est le prix à acquitter pour un niveau d'utilisation de l'infrastructure égal à x_i. Il détermine ces deux couples de façon à optimiser son profit. À cet objectif, s'ajoutent des contraintes liées au comportement des agents.

Les contraintes de participation assurent que les utilisateurs préfèrent utiliser l'infrastructure plutôt que de ne pas le faire :

$$u(x_1, t_1) - r_1 \geqslant 0$$

$$u(x_2, t_2) - r_2 \geqslant 0$$

Les deux contraintes suivantes sont incitatives. Elles évitent l'arbitrage personnel. Chaque type d'agent est mieux avec son niveau d'utilisation de l'infrastructure qui lui est destiné plutôt qu'avec celui pour l'autre type :

$$u(x_1, t_1) - r_1 \geqslant u(x_2, t_1) - r_2$$

$$u(x_2, t_2) - r_2 \geqslant u(x_1, t_2) - r_1$$

Le but de l'opérateur étant de pratiquer les prix les plus élevés possibles, il va chercher à saturer deux de ces quatre contraintes. Ainsi, pour les agents de type t_1 le prix est égal à leur disposition à payer maximale :

$$r_1 = u(x_1, t_1) \tag{1.2}$$

Pour les autres, le prix garantit qu'ils consommeront x_2 et non x_1 :

$$r_2 = u(x_2, t_2) - u(x_1, t_2) + u(x_1, t_1) \tag{1.3}$$

La fonction de profit de l'opérateur s'écrit donc :

$$\Pi(x_1, x_2) = [u(x_1, t_1) - cx_1]f_1 + [u(x_2, t_2) - u(x_1, t_2) + u(x_1, t_1) - cx_2]f_2$$

Le programme de l'opérateur est alors de maximiser ce profit par rapport aux deux niveaux d'utilisation de l'infrastructure. Les deux conditions du premier ordre sont :

$$\frac{\partial u(x_1, t_1)}{\partial x_1} = c + \left[\frac{\partial u(x_1, t_2)}{\partial x_1} - \frac{\partial u(x_1, t_1)}{\partial x_1}\right]\frac{f_2}{f_1} \tag{1.4}$$

$$\frac{\partial u(x_2, t_2)}{\partial x_2} = c \tag{1.5}$$

Les couples (r_1, x_1) et (r_2, x_2) se déduisent à partir de ces conditions du premier ordre (1.4)-(1.5) et des contraintes saturées (1.2)-(1.3).

Outre la discrimination par les quantités, il existe la discrimination par la qualité. Le modèle est le même que le précédent, sauf que x_i désigne le niveau de qualité. L'opérateur offre un éventail de qualités à des utilisateurs dont les goûts sont différents. Il modifie le niveau de qualité de ses services, afin de segmenter le marché. Le bien acheté est associé à un autre bien, « *un autre mal* » d'après TIROLE (1989) [20], puisqu'il s'agit de désagréments, tel que le temps d'attente ou une moindre qualité. Le comportement des agents vis-à-vis de ces inconvénients est perçu comme un signal sur leur propension à payer. Cela revient à dire que plus un agent est disposé à payer, moins il tolère un bien altéré.

L'offre de plusieurs classes de confort par des transporteurs est un exemple d'une discrimination tarifaire fondée sur plusieurs combinaisons prix-qualité.

1.2.3.2 Optimalité et surplus

TIROLE (1989) étudie la discrimination au second ordre, d'abord avec un tarif non linéaire binôme. Il montre que cette tarification domine au sens de PARETO la tarification monopolistique uniforme. L'opérateur peut réduire le prix marginal en-dessous du prix de monopole et récupérer les profits perdus avec la partie fixe.

Cependant, un tarif non linéaire par rapport aux quantités ou à la qualité est socialement sous-optimal. KATZ (1983) [9] montre qu'en général la discrimination au second degré conduit à des niveaux d'utilisation de l'infrastructure et de bien-être inférieurs aux niveaux de l'optimum collectif. Un opérateur discriminant accroît le bien-être par rapport au monopole non discriminant, mais n'élimine pas toute la perte résultant de la position de monopole.

L'exemple de la discrimination tarifaire par les quantités avec une offre de couples prix-quantité illustre cette situation. Avec la quantité qui leur est destinée, les utilisateurs à faible demande ont une disposition marginale à payer supérieure au coût marginal. Leur niveau d'utilisation de l'infrastructure est donc inférieur à celui socialement optimal.

Comme pour la discrimination par les quantités, seuls les utilisateurs avec une forte valorisation de la qualité vont acheter celle socialement optimale.

En revanche, les utilisateurs à forte demande achètent une quantité ou une qualité optimale. MIRRLEES (1971) [12] parle d'« *absence de distorsion au sommet* ». Selon la condition de SPENCE-MIRRLEES qui assure la monotonicité des quantités échangées avec le type, les agents de type t_2 souffrent plus d'une diminution de leur niveau d'utilisation de l'infrastructure. L'opérateur réduit donc la quantité consommée par les autres. Les utilisateurs de type t_2 sont alors moins tentés de s'intéresser au panier $(x_1; r_1)$. De plus, ils obtiennent une rente qualifiée d'informationnelle, alors que les utilisateurs du type t_1 ont un surplus nul. Cette rente incite les agents disposés à dissimuler leur type, à ne pas le cacher, au contraire à le révéler.

Cette situation tient à l'asymétrie d'information entre l'opérateur en monopole et les utilisateurs de l'infrastructure. Les quantités ou les qualités socialement optimales pour chaque type seraient obtenues si l'information était complète. Mais la mise en place du prix de premier rang en information incomplète n'aboutit pas aux quantités ou aux qualités de premier rang. C'est pourquoi, même l'optimisation du bien-être social en présence d'asymétrie d'information conduit à une solution de second rang, la même que celle qui maximise le profit.

Ce type de discrimination est largement employé par les opérateurs du secteur public. Les clients d'EDF et de France Télécom bénéficient d'une tarification binôme. Ils payent un abonnement chaque mois, augmenté d'une partie variable qui dépend du niveau d'utilisation du service, exprimé en kilowatt pour l'un et en temps de communication pour l'autre.

La tarification à la priorité (MARCHAND (1974) [11]) est également un moyen de discriminer au second degré. Ce mécanisme de prix sera présenté plus en détail à la section 2.2.3.2.

1.2.4 Le choix de la discrimination tarifaire

Le choix entre la discrimination liée à un signal exogène aux agents et celle liée au choix des agents se fait en fonction de l'objectif de l'opérateur ou de celui qui lui est assigné. Une amélioration du bien-être social peut être obtenue de plusieurs manières, selon que la discrimination tarifaire est du second ou du troisième degré. Celle au second degré accroît le bien-être grâce à une meilleure répartition de la demande, le plus souvent dans le temps. Les effets positifs de la discrimination au troisième degré sur le bien-être social sont dus à une plus grande accessibilité des agents à l'infrastructure, en raison de la baisse du prix pour certains utilisateurs, obtenus grâce à une hausse du prix pour d'autres, ceux ayant une plus forte disponibilité à payer.

L'objectif d'un opérateur privé portera plutôt sur la maximisation de son profit. Ce sont alors les informations dont il disposera sur les utilisateurs et la corrélation des caractéristiques de ces utilisateurs avec leur disponibilité à payer pour l'infrastructure, qui détermineront son choix pour une discrimination au second ou au troisième degré.

Ces deux modes de discrimination ne sont pas exclusifs l'un de l'autre. Il est possible de les combiner et ce choix est fréquemment fait dans le secteur des transports. La discri-

mination tarifaire sert, soit à accroître le profit de la société de transport, soit à augmenter le surplus social. Étant donné qu'il n'est pas évident de distinguer les voyageurs ayant une forte disponibilité à payer de ceux ayant une faible disponibilité à payer, les compagnies ont recours au *Yield Management*, une combinaison de la discrimination au second degré et de celle au troisième degré. De cette manière, les compagnies souhaitent se rapprocher de la discrimination parfaite.

Ainsi, pour distinguer un voyage d'affaires d'un voyage d'agrément, les entreprises de transport proposent différentes classes dont le confort et la qualité de service augmentent avec le prix. Les conditions de modification et d'annulation des billets servent aussi à discriminer. Plus ces conditions sont flexibles, le cas extrême permettant de modifier ou d'annuler sans frais, plus le prix est élevé. Les compagnies pratiquent aussi les tarifs de pointe : les billets en période chargée sont vendus plus cher que pour les vols en période creuse. Cette discrimination en fonction du moment du voyage n'est pas la seule. Les compagnies se servent également du fait qu'un passager passe un samedi soir ou plusieurs jours sur son lieu de destination comme signal, signifiant que le voyage est d'agrément. De cette manière, les passagers révèlent une faible disponibilité à payer et bénéficient de prix plus bas. L'âge est aussi une caractéristique des voyageurs corrélée avec leur propension à payer.

Comme nous l'avons vu plus haut, l'ensemble de ces discriminations conviennent aussi bien à un opérateur privé qu'à un opérateur public.

Conclusion

La figure (1.2) schématise la structure des prix : elle présente les déterminants fondamentaux des prix et leurs interactions.

Les interactions entre les coûts et les prix peuvent être de deux ordres. D'une part, la régulation « *cost-of-service* » repose sur la couverture totale des coûts. Des dépenses liées à l'infrastructure sont d'abord engagées et les prix sont ensuite déterminés de manière à ce que les recettes compensent les coûts. Par conséquent, c'est à partir des coûts que les prix sont établis, comme le signale le sens de la flèche sur la figure (1.2).

D'autre part, la régulation « *price-cap* » repose sur un plafond mis à l'évolution des prix. Les prix sont d'abord calculés en fonction du taux de progression autorisé et l'opérateur doit ensuite engager des dépenses qui ne dépassent pas les recettes anticipées. Par conséquent, c'est à partir des prix que le niveau des coûts est décidé, comme le signale le sens de la flèche sur la figure (1.2).

Les interactions entre la demande et les prix peuvent être de deux ordres. D'une part, la discrimination tarifaire au second degré laisse les utilisateurs choisir une combinaison associant au prix un niveau d'utilisation de l'infrastructure ou un niveau de qualité du service. À travers les différents prix proposés, l'objectif de l'opérateur est d'influencer la demande, par exemple de manière à l'étaler dans le temps lorsque la capacité est atteinte. Ainsi, en proposant différents niveaux de prix variant avec la quantité ou la qualité, l'opérateur s'attend à une modification de la demande des utilisateurs, comme le signale le sens de la flèche sur la figure (1.2).

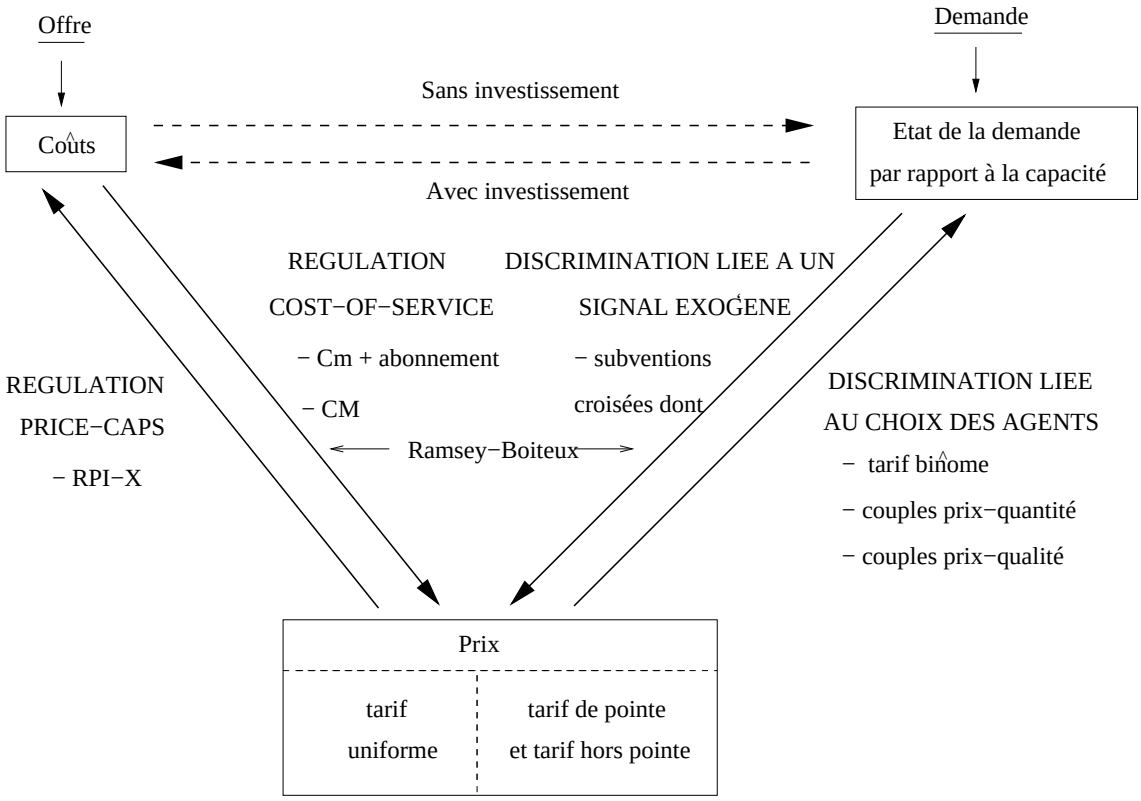

FIG. 1.2 – *Structure des prix.*

D'autre part, la discrimination tarifaire au troisième degré repose sur une différenciation des prix entre les utilisateurs de l'infrastructure sur la base de caractéristiques exogènes différentes entre ces agents. Les prix discriminants sont établis dans un souci d'accroître le bien-être, en rendant l'infrastructure accessible à un plus grand nombre d'utilisateurs, ou de maximiser le profit. C'est donc à partir de l'état de la demande, facilement segmentable et faible de la part de certaines catégories d'utilisateurs, que les prix sont fixés, comme le signale le sens de la flèche sur la figure (1.2).

L'infrastructure est mise en œuvre pour satisfaire un certain niveau de demande. Lorsque la demande est limitée par la contrainte de capacité, la question d'un investissement pour augmenter la capacité se pose. Quelle que soit la décision par rapport à l'investissement, celle-ci influence la structure des prix. La décision d'investir a des conséquences sur les coûts, puisque le montant de l'investissement est répercuté dans les coûts ; celle de ne pas investir a des conséquences sur l'état de la demande, car la pression de la demande sur la capacité est alors plus forte. Les prix devront prendre en compte soit le coût de l'investissement, soit l'existence d'une capacité insuffisante.

Chapitre 2

Théories de la tarification de pointe pour une infrastructure

Introduction

Le gestionnaire d'une infrastructure est face à un dilemme lorsqu'il doit déterminer le niveau de sa capacité. Soit cette capacité est plus faible que le niveau maximal de la demande : l'opérateur ne sera alors pas en mesure de satisfaire à tout moment la demande. Soit cette capacité est suffisante pour satisfaire la demande de pointe : il aura alors d'importantes capacités excédentaires pendant les périodes creuses. Une tarification de pointe peut contribuer à atténuer ce problème.

Sa justification s'appuie sur plusieurs arguments. Le premier, le plus communément admis, repose sur les coûts supportés par l'opérateur. Lorsque la capacité d'un service est dimensionnée de façon à ce qu'il puisse répondre à une forte demande, l'opérateur supporte un coût supplémentaire, celui de l'installation de cette capacité plus importante. Ce sur-coût dû à la capacité peut être inclus dans le calcul des prix.

La tarification de pointe se justifie aussi par la volonté d'opérer un certain rééquilibrage des demandes sur les différentes périodes. De ce point de vue, l'objectif des prix de pointe est de déplacer la demande des périodes de pointe vers les périodes creuses.

Une situation de surcapacité peut également être à l'origine d'une tarification de pointe. La demande exerce un effet externe négatif sur elle-même : une augmentation de la demande renforce la congestion et conduit à un service de moindre qualité. Une tarification de pointe ferait prendre conscience aux agents du surcoût qu'ils font supporter aux autres. Dans le souci d'une allocation efficace, un effet externe négatif se traduira par un coût pour celui qui le génère. Chaque agent doit subir la somme des effets qu'il fait ressentir aux autres ; un coût supplémentaire en période de pointe, égal au coût social, rétablirait la situation.

Une tarification de pointe peut donc se justifier par trois éléments principaux : le coût additionnel de la capacité, un mécanisme incitatif et l'internalisation des effets externes.

Mais ce choix ne se fait pas *a priori*. La solution optimale dépend essentiellement de la demande, selon qu'elle est déterministe ou aléatoire. Les deux sections de ce chapitre reposeront donc sur ces deux cas de demande.

2.1 Cas de la demande déterministe

Notre première approche de la théorie de la tarification de pointe se place dans le cadre d'une demande déterministe. Dans ce cas, la fonction de demande est connue pour chaque période. Il n'existe pas d'incertitude quant au niveau de la demande : l'opérateur sait quelle sera la quantité de service demandée pour chaque niveau de prix possible.

L'idée courante selon laquelle les prix doivent refléter les coûts, et qu'ainsi chaque utilisateur doit payer un prix en relation avec les coûts qu'il a induits, est applicable dans ce contexte. Face à une telle demande, il est facile pour l'opérateur de faire des prix le reflet des coûts de fonctionnement et d'investissement de l'infrastructure. Cependant, selon différentes configurations de niveaux de coûts et d'écart entre les demandes, de pointe et de hors pointe, les prix diffèrent.

En résolvant le programme de maximisation du surplus social, nous verrons quelles recommandations préconiser en matière de définition de prix. Mais auparavant, étant donné que ces prix sont fixés au regard des coûts, il est important de s'intéresser aux coûts d'un opérateur. De nombreux coûts peuvent être identifiés.

2.1.1 Identification des coûts

Le coût global d'une infrastructure supporté par un opérateur ou par la société peut être divisé en de nombreux coûts. Certains de ces coûts ont une partie de leur assiette commune. D'autres ne relèvent pas nécessairement de la sphère marchande ou ne correspondent pas strictement aux coûts comptables. Nous donnerons une définition de la plupart de ces coûts.

Cependant, tous n'ont pas un rôle à jouer dans la détermination des prix. Le coût pertinent pour établir une tarification efficace est le coût marginal. Mais la question elle-même du périmètre du coût marginal se pose.

2.1.1.1 Les différentes catégories de coûts

CURIEN (2000) [8] associe par paires toute une série de coûts. Nous reprenons ici cette classification, afin d'identifier les nombreux coûts existants :

- coûts fixes / coûts variables : la fourniture d'un bien engendre, d'une part, des coûts fixes, indépendants du volume de la production, d'autre part, des coûts variables, croissants en fonction de ce volume ;
- coûts communs / coûts séparables : pour assurer la fourniture de plusieurs services une seule infrastructure peut suffire ; se pose alors la question de l'imputation de ces coûts. Lorsqu'ils sont communs ou joints, on ne sait pas quelle part de ces coûts incombe à

quel service ; si les coûts sont séparables, ils sont au contraire clairement identifiables et peuvent être attribués à chaque service ;

- coûts marginaux / coûts moyens : en s'intéressant au coût d'une unité de production, on peut calculer soit le coût moyen correspondant en moyenne à ce que coûte la production d'une unité, soit le coût marginal portant sur le coût de la dernière unité produite ;
- coûts incrémentaux / coûts isolés : le coût incrémental d'un service désigne la part dans le coût total consacré uniquement à ce service, le coût isolé serait le coût du service s'il devait être rendu seul. Au sein d'un monopole naturel, le coût incrémental de fourniture d'un service est inférieur à son coût de fourniture isolé, car en étant servi avec d'autres services, on évite la duplication d'une partie des coûts ;
- coûts historiques / coûts de renouvellement : on peut considérer que la valeur de l'infrastructure correspond soit aux coûts historiques, supportés par l'opérateur au moment de l'installation, avec une réévaluation due à l'inflation, soit à ce que cela coûterait de remplacer l'infrastructure ;
- coûts recouvrables / coûts non recouvrables : les coûts recouvrables sont les coûts récupérables par un opérateur lorsqu'il abandonne son activité : on parle alors d'un marché parfaitement « contestable », lorsque des producteurs peuvent sortir du marché en recouvrant leurs coûts d'entrée ;
- coûts comptables / coûts économiques : le coût comptable de l'opérateur correspond à la valeur de ses dépenses monétaires, incluant la dépréciation de l'infrastructure et d'autres transactions comptables. Les coûts économiques sont beaucoup plus larges : ils incluent des coûts que l'opérateur n'a pas eu à supporter directement et qui ne relèvent pas de la sphère marchande, tels que le coût d'opportunité de son activité (coût égal au rendement de la somme investie dans l'infrastructure, si elle l'était dans une autre activité), ou tels que les coûts externes liés par exemple aux nuisances environnementales ou à la congestion.

2.1.1.2 La décomposition du coût marginal

Dès lors que l'on cherche à maximiser le bien-être collectif, l'économie publique enseigne que le bien ou le service doit être vendu à un prix égal au coût marginal. Cette solution est l'optimum de premier rang : elle assure qu'avec ce prix le surplus collectif le plus élevé possible est atteint.

Bien que la notion de coût marginal semble correspondre à un coût bien spécifique, son périmètre reste à définir.

Le modèle présenté à la section suivante (STEINER, 1957 [19]) s'appuie sur deux types de coûts ; ils sont supposés être linéaires. Le premier coût marginal est celui de la production, également dit coût de fonctionnement. Il concerne les charges liées à son exploitation et à son entretien. Ce coût marginal, portant sur des dépenses effectuées régulièrement, privilégie le court terme ; il est désigné par la suite avec la lettre b.

Le second coût marginal est celui de la capacité. Il correspond au coût par jour pour disposer d'une nouvelle unité de capacité ; la notation pour ce coût sera β. En ajoutant ce coût au précédent, on obtient un coût marginal de long terme. Cette notion est employée pour

indiquer qu'il s'agit d'une optimisation dans la durée, afin que le choix des investissements soit optimal.

2.1.2 Justification de la tarification de pointe par les coûts de l'opérateur

Nous présentons dans cette section les résultats de STEINER (1957) qui s'intéresse aux prix qui maximisent le bien-être collectif dans une situation de pointe avec des demandes déterministes.

On distingue deux périodes de durée fixée, chacune caractérisée par une fonction de demande différente : $D_1(P_1)$ et $D_2(P_2)$, la courbe associée à la seconde fonction étant toujours située au dessus de la courbe de la première. Les demandes sont supposées indépendantes : le prix appliqué à une période n'a aucun effet sur la quantité demandée à l'autre période.

L'hypothèse sur les coûts est très forte. Comme nous l'avons vu à la section 2.1.1.2), les coûts sont linéaires. Cela signifie que les coûts marginaux opérationnels et de capacité sont constants. Servir la demande d'une unité au cours d'une période, coûtera b si la capacité existe déjà pour cette unité ou coûtera $b + \beta$ si une unité de capacité additionnelle doit être installée. Une fois qu'une unité de capacité est mise en place, elle peut servir les demandes des deux périodes.

La solution au programme d'optimisation aboutit à des prix qui reflètent les coûts. On différencie, cependant, deux cas, selon les niveaux de coûts et l'écart entre la demande de pointe et celle de hors-pointe. Dans le premier cas, les deux demandes sont suffisamment éloignées. La demande de pointe ne sera pas attirée par un prix plus bas en période creuse : elle restera stable. Dans le second cas, les deux demandes sont proches l'une de l'autre, et si le coût marginal de fonctionnement de l'infrastructure est très faible et si celui de l'accroissement de capacité est très élevé, le bas niveau du prix en période creuse va attirer « trop » d'utilisateurs à cette période. La demande la plus faible risque alors de devenir la demande de pointe. Il faut pallier cette instabilité en modifiant la règle de prix.

2.1.2.1 Avec une pointe stable

Étant donné que nous sommes dans une situation de demande déterministe, le niveau de capacité optimal, K, s'égalise au niveau de la demande la plus élevée :

$$K = \max(X_1, X_2) \tag{2.1}$$

Il est inutile d'avoir $K > X_1$ et $K > X_2$, cette capacité excédentaire ne servirait pas et serait coûteuse à installer. Avec l'égalité (2.1), les coûts de congestion sont par conséquent nuls car la demande est toujours satisfaite.

Dans le cas qualifié de « pointe stable », le résultat de l'optimisation conduit à des quantités échangées différentes aux deux périodes :

$$X_1^S < X_2^S = K^S$$

Dans cette configuration, l'opérateur est en mesure de faire payer aux utilisateurs les coûts exacts qu'ils induisent. Ainsi les utilisateurs de la période creuse, qui ne nécessitent pas un

accroissement de capacité, s'acquittent simplement du coût marginal opérationnel :

$$P_1^S = b^S$$

Les utilisateurs de la période de pointe, à l'origine d'une capacité plus importante, payent pour cette capacité supplémentaire :

$$P_2^S = b^S + \beta^S$$

La figure (2.1) illustre cette situation. La demande de pointe est beaucoup plus importante que la demande en période creuse : pour un prix donné, la quantité demandée est plus élevée. Les prix P_1^S et P_2^S donnent les quantités échangées X_1^S et X_2^S et le niveau de capacité K^S est fixé égal à X_2^S.

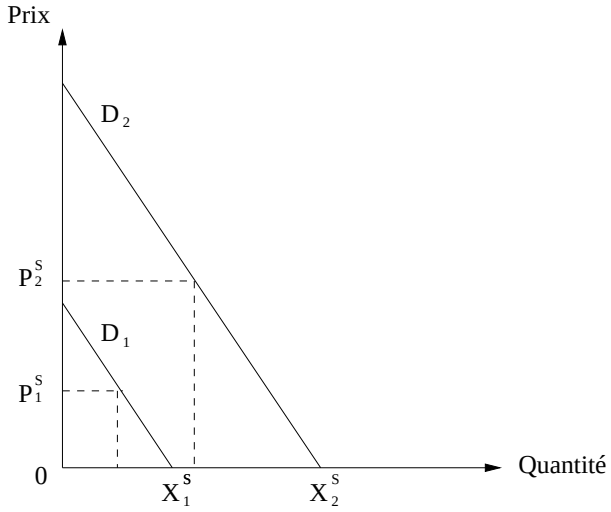

FIG. 2.1 – *Cas d'une pointe stable*

Si l'égalisation des prix aux coûts marginaux ne permet pas de couvrir les coûts fixes, une tarification RAMSEY-BOITEUX est possible. Elle définit pour les utilisateurs de la période creuse, un taux de marge par rapport au coût marginal b^S inversement proportionnel à l'élasticité-prix de la demande de cette période :

$$\frac{P_1'^S - b^S}{P_1'^S} = \frac{\lambda^S}{1+\lambda^S} \frac{1}{\mid \epsilon_1^S \mid}$$

Le taux de marge appliqué aux utilisateurs de la période de pointe est déterminé par rapport aux coûts marginaux opérationnel et de capacité. Il est également proportionnel à l'élasticité-prix de la demande de cette période :

$$\frac{P_2'^S - (b^S + \beta^S)}{P_2'^S} = \frac{\lambda^S}{1+\lambda^S} \frac{1}{\mid \epsilon_2^S \mid}$$

2.1.2.2 Avec une pointe instable

Néanmoins, avec la règle de prix précédente, il existe un risque de déplacement de la pointe. Ce fut la conséquence en Allemagne de l'instauration d'un tarif de nuit pour le téléphone :la baisse du prix des communications à partir de 22 heures eut pour conséquence d'accroître la demande de 22 heures à 23 heures. Ce système fut donc abandonné en 1981 (BÖS, 1985 [5]).

En effet, dans certaines circonstances de coûts et de demandes, la situation n'est pas stable. Prenons l'exemple d'un coût marginal opérationnel plus faible $(b^I < b^S)$, un coût marginal de la capacité plus élevé $(\beta^I > \beta^S)$ et des courbes de demande rapprochées. Cette situation est représentée sur la figure (2.2). On s'aperçoit qu'avec les prix P_1^S et P_2^S, la demande de la période creuse est au-dessus de celle de la période de pointe et excède donc la capacité. D'autres prix doivent être envisagés.

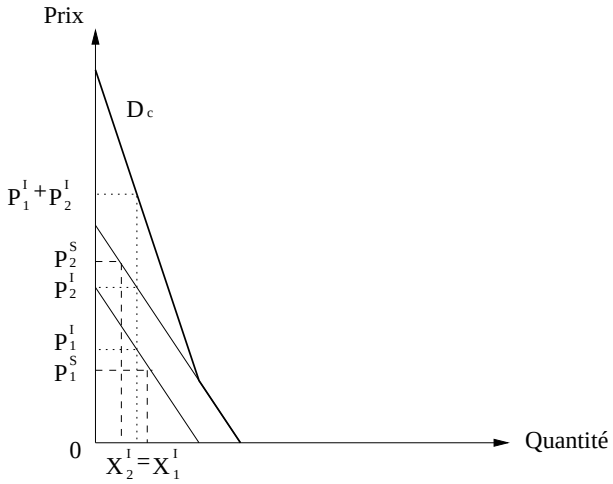

FIG. 2.2 – *Cas d'une pointe instable.*

Ce cas de « pointe instable » est également étudié par STEINER (1957). La maximisation du bien-être social dans cette situation aboutit à une nouvelle règle de prix.

La situation optimale consiste à additionner les deux courbes de demande pour n'en avoir plus qu'une :

$$X_c^I(P_c^I) = X_1^I(P_1^I) + X_2^I(P_2^I)$$

L'intersection entre cette nouvelle droite et un prix égal à $2 \times b^I + \beta^I$ donne la capacité optimale K^I. On retrouve une égalité qui prévalait aussi dans la situation précédente :

$$P_1^I + P_2^I = 2 \times b^I + \beta^I$$

Mais la répartition des coûts se fait différemment car :

$$P_1^I > b^I \text{ et } P_2^I < b^I + \beta^I \tag{2.2}$$

La différence de prix n'est plus égale à la différence de coûts. Au sens de PHLIPS (section 1.2), cette situation s'apparente donc à de la discrimination tarifaire.

Lorsqu'un problème de couverture des coûts fixes se pose, la tarification RAMSEY-BOITEUX dans ce cas se traduit par :

$$\frac{P_1'^I + P_2'^I - (2 \times b^I + \beta^I)}{P_1'^I + P_2'^I} = \frac{\lambda^I}{1 + \lambda^I} \frac{1}{\mid \epsilon_c^I \mid}$$

où ϵ_c^I est l'élasticité-prix de la demande globale D_c^I.

Dans une situation de pointe instable, les prix doivent donc être tels que la demande de pointe s'égalise à la demande hors pointe. Ce résultat signifie que les utilisateurs de la période creuse payent une partie des coûts de la capacité supplémentaire.

La validation de ces résultats, dans les deux situations, reste soumise à l'hypothèse de linéarité des coûts.

2.1.3 Régulation par les prix et tarification de pointe

La tarification de pointe est tout à fait compatible avec une régulation « cost-of-service ». Le recouvrement des coûts peut se faire en séparant les coûts de la période de pointe de ceux de la période hors pointe.

En revanche, une tarification de pointe est plus difficile à concilier avec une régulation « price-cap ». Ce type de régulation n'est pas flexible et ne permet pas de changer brusquement le plafond de prix pour une courte période.

De plus, les motivations de chacun de ces deux principes de prix vont à l'encontre les unes des autres. L'intérêt d'instaurer des prix de pointe, liés aux coûts, est de faire refléter à travers les prix l'ensemble des coûts supplémentaires qui surviennent en période de pointe. Les prix de pointe peuvent ainsi être sensiblement plus élevés que les prix hors pointe. À l'inverse, la régulation « price-cap » repose sur la mise en place d'un plafond à l'évolution des prix afin de limiter les dépenses de l'opérateur. Il est donc difficilement envisageable d'autoriser un opérateur à pratiquer des prix de pointe reflétant les coûts. Il serait alors en mesure de dégager une forte rente et ne serait plus soumis à une maîtrise de ses coûts.

Une régulation « price-cap » limite l'utilisation des prix comme moyen de faire face à une infrastructure saturée, puisque son intérêt est de conserver des prix bas. La compatibilité entre la tarification de pointe et le « price-cap » fait partie des domaines actuels de recherche en matière de régulation (BRUNEKREEFT, 2000 [4]).

2.2 Cas de la demande aléatoire

Lorsque la demande n'est plus déterministe, mais aléatoire, il n'est plus possible d'égaliser en permanence l'offre à la demande la plus forte.

La question du dimensionnement de l'infrastructure est plus complexe dans ce contexte. Si la capacité est fixée de manière à couvrir la demande en toute circonstance, même dans

les cas où la demande est considérablement élevée, cas de très faible probabilité, cette offre risque d'être très coûteuse et inefficace du fait de l'inutilisation de l'infrastructure au cours d'autres périodes. Avec une capacité qui ne permet pas de satisfaire la demande à tout instant, une partie de la demande risque d'être « rationnée », c'est-à-dire non servie au moment où elle s'exprime. Se pose alors la question de l'utilisation optimale de l'infrastructure existante.

Ces deux interrogations, « Quelle est la capacité optimale ? » et « Comment allouer efficacement cette capacité ?», peuvent être traitées à partir de la tarification de pointe. Après avoir discuté de la taille optimale de l'infrastructure, nous répondrons à la seconde question. Des prix de pointe sont un moyen d'allouer efficacement l'offre.

2.2.1 La taille de l'infrastructure

Les choix relatifs à la taille de l'infrastructure sont guidés par le coût marginal de la capacité et son bénéfice marginal. Ce bénéfice mesure en termes monétaires de combien d'unités le bien-être social s'accroît à la suite d'une augmentation de la capacité de l'infrastructure.

Il peut aussi et surtout être vu comme un coût évité. Un rationnement de la demande est source de congestion qui est à l'origine de coûts supplémentaires, notamment des coûts de retard. Un accroissement de la taille de l'infrastructure pour résoudre le problème de capacité insuffisante aurait pour conséquence une réduction de ces coûts de congestion, ce qui correspondrait à l'avantage tiré de l'investissement. La réduction des coûts de congestion est à l'origine d'un bénéfice positif.

Cependant, nous verrons qu'il est parfois optimal de rationner la demande.

2.2.1.1 La règle de capacité optimale

L'investissement dans une infrastructure est coûteux. Une capacité trop élevée par rapport à la demande risque d'être une source de gaspillage, en raison de l'inutilité d'une partie de l'infrastructure. Cependant, les perspectives de croissance de la demande doivent être prises en compte. En raison des délais liés à l'augmentation de la capacité, il peut être préférable qu'une partie de l'infrastructure soit inutilisée pendant un certain temps plutôt que d'attendre que l'infrastructure soit saturée pour lancer un nouvel investissement.

En effet, les conséquences d'une capacité insuffisante peuvent également être coûteuses, les principaux coûts d'une telle situation étant liés à la congestion. Néanmoins, cela ne signifie pas qu'il faut satisfaire la demande à tout moment et éviter la congestion « à tout prix ». Le dimensionnement de la capacité doit se faire à un niveau tel que l'utilité de le relever à la marge équilibre exactement le coût de développement.

Si le bénéfice d'une augmentation de la capacité est supérieur au coût marginal, alors il est nécessaire d'accroître la capacité afin d'augmenter le bien-être social. Le bénéfice généré par l'accroissement de la capacité fera plus que compenser le coût de cette hausse. À l'inverse, si le bénéfice marginal est inférieur au coût marginal, alors il faut définir un niveau de capacité plus faible. Le bénéfice procuré par l'accroissement de capacité ne permet pas de couvrir le coût supplémentaire. L'égalité entre le coût marginal de l'infrastructure et son bénéfice marginal assure donc une capacité optimale.

2.2.1.2 Le rationnement de la demande

Nous venons de voir que l'optimisation de la capacité n'aboutit pas nécessairement à satisfaire toute la demande. Il est préférable lorsque le bénéfice d'une augmentation de capacité est inférieur à son coût de rationner la demande. Dans un modèle de demande de pointe, BÖS (1985) s'intéresse à l'arbitrage entre le rationnement par les prix et le rationnement par les quantités.

La fonction de demande retenue se caractérise par une dépendance au temps, au sein d'une même période. Ainsi, pour une période e appartenant à l'ensemble de toutes les périodes E, la quantité demandée par unité de temps (dans une période) s'écrit :

$$x_e(p_e, t_e) = x_e(p_e) + \tau(t_e)$$

où le prix p_e est le même pour toutes les unités de temps t_e de la période e, ce qui signifie que les élasticités-prix croisées sont nulles. La fonction est séparable : la sensibilité de la demande au prix est indépendante de la période et la sensibilité de la demande à la période est indépendante du prix.

Du côté de l'offre, la quantité disponible par unité de temps est indépendante du temps : elle est constante au cours d'une période, mais peut varier d'une période à une autre.

La quantité vendue correspond au minimum entre la quantité demandée et celle offerte. S'il y a un excès de demande, celle-ci doit être rationnée. Ce rationnement peut soit être aléatoire, soit reposer sur les disponibilités à payer des utilisateurs. Cette distinction se formule comme la différence entre le rationnement par les quantités et celui par les prix. Un rationnement de la demande, se traduit par une diminution du bien-être social. La manière dont ce « *welfare* » est affecté dépend de l'importance que l'opérateur accorde au fait que la demande ne puisse être entièrement satisfaite. La maximisation du bien-être social, sous diverses contraintes, notamment celle de l'équilibre budgétaire ou encore un niveau suffisant de la fiabilité de l'offre, aboutissent à certaines conditions d'optimalité.

Tout d'abord, dès qu'il y a des fluctuations de demande pendant une période donnée e, il est optimal d'avoir un excès de demande. La demande doit donc être rationnée.

Ensuite, le mode de rationnement repose sur la sensibilité de l'opérateur au rationnement. S'il considère que rationner la demande revient à diminuer nettement le bien-être, alors il préférera gérer cette situation à travers la pratique de prix élevés. Au contraire, si l'opérateur accorde peu d'importance aux effets sur les utilisateurs d'un rationnement, il choisira des prix faibles, au voisinage du coût marginal, et organisera un rationnement par les quantités.

Enfin, se pose la question de savoir comment être assuré que le prix hors pointe n'excèdera pas le prix de pointe. Plusieurs conditions doivent être réunies :

- l'exigence budgétaire fixée par le régulateur à l'opérateur doit être assez forte ;
- l'opérateur doit être suffisamment sensible à l'excès de demande et au rationnement ;
- l'élasticité-prix de la demande de pointe doit être proche de celle de la demande des périodes creuses ;
- le coût marginal de pointe ne doit jamais être en dessous de celui de hors pointe.

2.2.2 Justification de la tarification de pointe par les coûts de congestion

L'existence de coûts de congestion peut justifier un tarif de pointe pour l'usage d'une infrastructure. Le coût marginal de congestion, dont il va être désormais question correspond au dommage causé par un utilisateur i aux autres utilisateurs lorsqu'il augmente à la marge son utilisation de l'infrastructure. Ce coût représente la contribution supplémentaire que tous les autres utilisateurs seraient prêts à payer pour que i réduise à la marge son utilisation de l'infrastructure.

Sans l'intervention d'une autorité extérieure, c'est coûts ne sont pas inclus dans le calcul économique alors qu'ils sont tout de même supportés par une partie des utilisateurs, lorsqu'ils ne disposent pas de l'infrastructure au moment où ils le souhaitent. L'idée des systèmes de prix qui vont être présentés est d'internaliser ces coûts, en incitant les agents à prendre en compte ces coûts de congestion à travers un péage correspondant aux effets externes qu'ils engendrent. Ces coûts étant nuls aux périodes creuses, ce mécanisme introduit une différence de prix entre les périodes chargées et celles qui le sont moins.

En matière de tarification de la congestion temporelle, les premiers apports théoriques ont concerné le transport sur route. Des modèles similaires ont ensuite été développés pour la congestion aéroportuaire, les infrastructures devenant saturées. Cependant, à travers une comparaison air-route, BRUECKNER (2002) [3] met en garde contre l'analogie souvent faite pour traiter ces deux types de congestion.

2.2.2.1 Modèle de congestion routière

VICKREY (1969) [21], puis ARNOTT, DE PALMA et LINDSEY (1993) [1] ont étudié une situation de congestion routière et suggéré la mise en œuvre d'un péage visant à résorber un « bouchon » et à rendre la circulation plus fluide. Nous présentons ici le second modèle qui s'appuie sur le premier.

Le modèle s'intéresse à une multitude d'agents qui souhaitent arriver au même endroit, au même moment et qui présentent les mêmes caractéristiques. Pour y parvenir, ils doivent franchir un passage rétréci, comme celui représenté sur la figure 2.3. Ce goulot d'étranglement a un débit fixe inférieur au nombre de personnes qui souhaitent passer à cet endroit. On observe donc la formation d'une file d'attente.

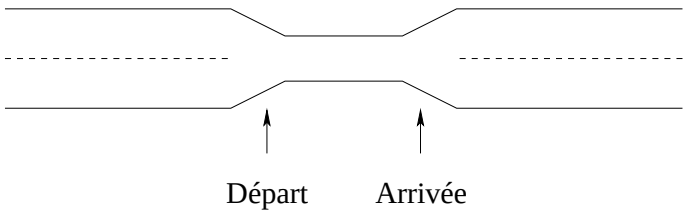

FIG. 2.3 – *Goulot d'étranglement.*

Le coût total du trajet se compose de deux éléments : un coût privé et un péage s'il y a lieu. Le coût privé est supposé linéaire par rapport au temps d'attente et à l'écart entre l'heure d'arrivée et l'heure souhaitée.

Chaque agent doit décider à quel moment il part de chez lui. Il fait ce choix en confrontant son temps de trajet, son retard ou son avance par rapport à l'heure d'arrivée souhaitée et le prix du péage. L'équilibre est atteint lorsqu'aucun agent ne peut réduire le coût de son trajet en changeant son heure de départ, les heures de départ des autres agents étant considérées comme données [1].

En l'absence de péage, l'équilibre est atteint lorsque, quelle que soit l'heure de départ des agents, ceux-ci ont des coûts privés identiques pour le trajet.

Nous savons qu'à l'équilibre, le coût doit être le même pour tous. Ainsi,

- le coût du premier à partir, celui qui n'aura pas de temps d'attente, mais qui sera le plus en avance,
- le coût du dernier à partir, celui qui n'attendra pas également, mais qui sera le plus en retard et
- le coût du seul qui arrivera à l'heure désirée, mais qui aura attendu le plus

devront être égaux.

Les différentes conditions d'équilibre nous permettent de calculer les bornes de l'intervalle de temps où il y a de la congestion $[t_q, t'_q]$, ainsi que le taux d'arrivée des agents en avance, sur l'intervalle $[t_q, \overline{t}]$, et celui des agents en retard, sur l'intervalle $[\overline{t}, t'_q]$.

La file d'attente évolue donc en fonction de deux taux d'arrivée. La figure (2.4) est une représentation graphique de l'évolution de la file d'attente et des heures de sortie de l'endroit congestionné. La distance 1, en t, mesure le nombre d'agents qui rejoignent le « bouchon »,

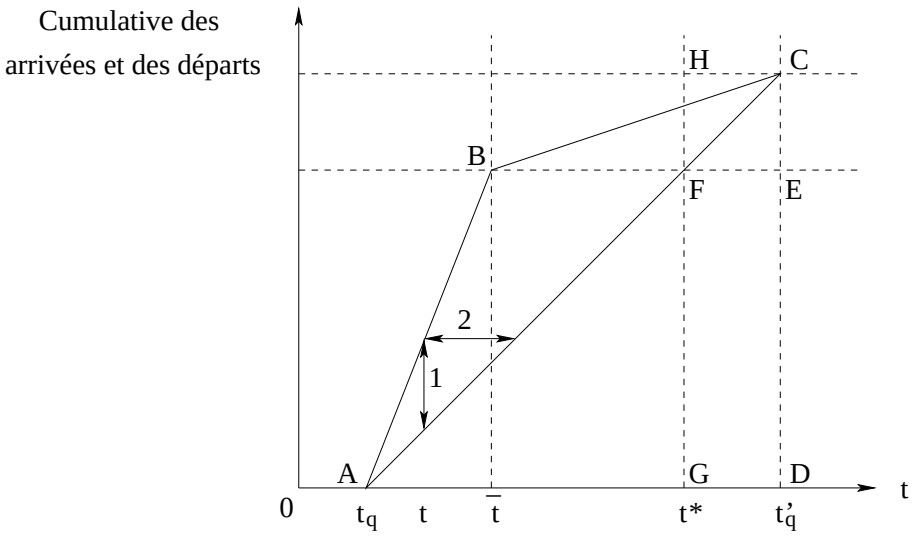

FIG. 2.4 – *Équilibre sans péage.*

c'est la longueur de la file d'attente. La distance 2 nous indique le temps d'attente des agents qui se présentent à l'entrée du goulot d'étranglement en t. La droite (AC) correspond aux sorties cumulées, il s'agit d'une droite car ces sorties sont continues et se produisent selon

1. On parle d'équilibre de NASH, avec pour variables stratégiques les heures de départ.

le taux de débit du goulot d'étranglement. Les segments $[AB]$ et $[BC]$ correspondent aux arrivées cumulées au goulot. La file d'attente est d'abord linéaire de t_q à $\overline{t}$. Puis, nous observons une cassure en $\overline{t}$, car c'est à cet instant que le temps d'attente est le plus long. C'est aussi l'heure de départ qui permet d'arriver à l'heure souhaitée t^*. Enfin, la file d'attente est toujours linéaire, mais avec une pente plus faible, de $\overline{t}$ à t'_q. Le segment $[AB]$ est plus pentu car les agents « s'accumulent » du fait du bouchon.

À partir de ce graphique, nous pouvons calculer différents coûts. Le temps de trajet correspond à l'aire du triangle (ABC). En multipliant cette quantité par le coût unitaire du temps d'attente, nous obtenons le coût total du temps de trajet d'équilibre.

L'arrivée décalée, par rapport à l'heure d'arrivée souhaitée t^*, se mesure, d'une part, avec les agents en avance, dont le temps d'avance correspond au triangle (AFG), et d'autre part, avec les agents en retard, dont le temps de retard est représenté par le triangle (CFH). En multipliant ces quantités respectivement par le coût unitaire d'une arrivée en avance et le coût unitaire d'une arrivée tardive, nous obtenons le coût total d'équilibre de l'arrivée décalée. Son montant est le même que celui du coût total du temps de trajet d'équilibre.

Le coût total d'équilibre du trajet est égal à la somme des deux coûts précédents.

À partir de ce coût total, pour l'ensemble de la société, nous pouvons calculer le coût total moyen et le coût total marginal. Le coût total moyen correspond au coût individuel supporté par chaque agent, en l'absence de péage. Le coût social marginal correspond au coût que chacun supporte à la suite de l'arrivée d'un nouvel agent dans la file d'attente. Il est égal à deux fois le coût total moyen.

En l'absence de péage, le coût pour chaque agent est égal à son coût privé, correspondant en moyenne au coût total moyen. Or, il existe une différence entre le coût social et le coût privé. Un nouvel entrant va supporter un coût égal au coût total moyen, alors qu'il va imposer à l'ensemble de la société le coût social marginal, deux fois supérieur au précédent. Un agent, qui souhaite traverser l'endroit congestionné, va « sous-estimer » l'impact de son arrivée dans la file sur les autres agents. Un mécanisme de péage, internalisant l'effet sur la société, peut rétablir la situation.

De plus, la formation de la file d'attente est à l'origine d'une perte sèche. Si le taux de départ était égal au débit du goulot d'étranglement dans l'intervalle de temps $[t_q, t'_q]$, il n'y aurait plus de file d'attente et le coût total du temps de trajet serait nul. Or, ce coût représente la moitié du coût total du trajet. Ainsi, une économie égale à la moitié du coût total du trajet serait possible sans modifier la distribution d'arrivée et le nombre de trajets.

La perte sèche aurait pu être économisée si la demande avait été uniformément répartie entre les moments t_q et t'_q. Cependant, cette distribution de l'arrivée des agents au goulot ne peut s'obtenir qu'avec des incitations. Ces incitations financières font que la perte sèche ne disparaît pas totalement, mais est récupérée par une autorité chargée du péage.

Pour qu'il soit efficace, le péage doit dépendre du temps. L'intérêt de ce péage serait d'éviter les temps d'attente, afin que le trafic soit « fluide ». L'optimum social décentralisé consiste donc à faire payer un droit de passage de façon à ce que la file d'attente disparaisse et donc que le coût lié au temps de trajet soit nul. Le coût total optimal est simplement égal

au coût de l'arrivée décalée à l'équilibre.

Les coûts totaux moyen et marginal optimaux représentent alors la moitié de leur montant d'équilibre. Le coût total marginal est donc toujours égal à deux fois le coût total moyen.

Le coût supporté par chaque agent doit correspondre au coût total marginal. Ainsi, celui qui arrive à l'heure, qui a donc un coût privé nul, doit s'acquitter d'un péage égal à deux fois le coût total moyen. En revanche, les deux agents qui passent le goulot en premier et en dernier subissent un coût dû à leur arrivée décalée déjà égal à deux fois le coût total moyen. Ils ne paient donc pas de péage. Le droit de passage est croissant pour une traversée entre t_q et t^*, puis décroissant entre t^* et t'_q.

En moyenne, le péage est égal au coût total moyen. Ce montant correspond à la différence entre le coût réel supporté par les agents et le coût social marginal, lorsqu'il n'existe pas de péage. En instaurant le péage, les agents sont face aux coûts qu'ils imposent aux autres, et de cette manière ils peuvent internaliser les conséquences de leur choix d'heure de passage.

Ainsi, en situation de congestion de pointe, une partie des coûts privés supportés par les agents peut être évitée. Les coûts liés au fait d'être en avance ou en retard sont eux incompressibles. Mais, l'instauration d'un péage permet de réduire les coûts liés à l'attente, dus à une distribution inefficace des heures de départ des agents dans le temps. Afin d'obtenir une répartition efficace de la demande dans le temps, le péage doit compenser les écarts de coûts entre les agents dont les heures de départ sont différentes. Ainsi, on peut éviter tout arbitrage. Un péage parfaitement discriminant, le péage optimal, est tel que son montant change à chaque instant, afin de respecter l'égalité des coûts totaux (coûts privés plus péage) entre les agents.

2.2.2.2 Modèle de congestion aéroportuaire

Le modèle de DANIEL (1995), très proche de celui de ARNOTT, DE PALMA et LINDSEY (1993), étudie la tarification liée aux problèmes de congestion de pointe dans les grands aéroports. Il part du constat que les taxes d'aéroport sont fondées sur le poids des avions et sont indépendantes du niveau de congestion. Or, il existe un coût social pour les atterrissages et les décollages, qui correspond aux coûts de retards supplémentaires. Ceux-ci sont pourtant indépendants du poids des avions, mais dépendants de la congestion de l'aéroport. Le facteur poids est donc un mauvais élément qui ne permet pas de choisir des heures de départ ou d'arrivée qui minimisent le coût social. Les atterrissages et les décollages sont prévus sur de très courtes périodes qui sont très demandées, alors que la situation optimale consisterait à étaler le trafic. L'objectif du modèle de DANIEL est d'améliorer l'utilisation de la capacité à travers la tarification.

Un modèle de tarification de la congestion de pointe, avec une demande stochastique semble convenir pour les aéroports servant de *hubs* à certaines compagnies. En effet, ces aéroports sont confrontés à une demande très fluctuante, et présentent des périodes où la pointe est très forte et où les files d'attente sont longues. Un arbitrage doit donc être fait entre

les retards, dus à la congestion, et le temps d'escale, élément caractéristique des *hubs*, qui augmente si la période de pointe est étalée.

À partir d'une telle situation, DANIEL (1995) calcule le coût marginal imposé par un atterrissage supplémentaire. Pour cela il commence par évaluer le coût pour un vol d'atterrir à la période t, alors que l'heure souhaitée est s_n. Ce coût est la somme des coûts d'escale, de chevauchement et de file d'attente. Avec leur organisation en *hub and spoke*, les compagnies réduisent leurs coûts d'exploitation et les coûts liés au décalage par rapport à l'heure à laquelle les passagers auraient souhaité voyager. La baisse de ces coûts est en fait compensée par des trajets rallongés, des problèmes de liaisons et des temps d'escale. Pour minimiser ces coûts supplémentaires, les compagnies constituent des « plages » d'arrivées et de départs. Il s'agit de faire décoller ou atterrir tous les vols à destination ou en provenance des villes *spoke* approximativement à la même heure, afin que toutes les combinaisons de correspondances soient possibles.

Le coût d'escale correspond au coût occasionnée par le temps passé au *hub* entre le moment où l'avion quitte la file des atterrissages et où il entre dans la file des décollages. Si tous les vols à destination du *hub* n'arrivent pas avant le premier départ, il existe un risque pour les passagers de rater leur correspondance, dont le coût est celui de chevauchement. Par ailleurs, les aéroports sont soumis à des contraintes de capacité. Si les heures d'atterrissage et de décollage ne sont pas respectées, les avions doivent entrer dans des files d'attente, dans les airs pour atterrir et sur les pistes pour décoller, le temps passé à attendre étant coûteux.

L'espérance de coût de départ d'un avion est la somme de ces trois coûts sur toutes les périodes t, pondérés par la probabilité d'atterrir en t. L'heure optimale de l'avion n, s_n, pour rejoindre la file des atterrissages s'obtient en optimisant le programme du planificateur social, c'est-à-dire en minimisant les coûts pour l'ensemble des compagnies. Or, elle ne correspond pas à la solution du programme de maximisation d'une seule compagnie. Une compagnie aérienne prise individuellement ne tient pas compte de l'effet de sa décision d'opérer un vol supplémentaire sur l'espérance de coût de départ de tous les avions. Elle considère que ce coût est exogène, alors qu'en opérant un vol n, la compagnie impose une hausse des coûts aux autres vols. Le régulateur peut obtenir un optimum décentralisé en introduisant une tarification de la congestion de façon à ce que cet effet externe soit « internalisé ».

Comme ARNOTT, DE PALMA et LINDSEY (1993), DANIEL (1995) recommande l'introduction d'une surtaxe égale au coût marginal d'une arrivée supplémentaire. En faisant supporter aux compagnies le coût marginal de la congestion qu'elles imposent, elles sont amenées à adopter des comportements optimaux.

2.2.2.3 Comparaison air-route

Un autre apport intéressant de la littérature à la question qui nous préoccupe est celui de BRUECKNER (2002). Comme beaucoup d'articles (ARNOTT, DE PALMA et LINDSEY, 1993, DANIEL, 1995, MORRISON, 1987) sur le transport aérien, l'auteur part du constat des importants retards qui existent autant aux États-Unis qu'en Europe. Face à l'insuffisance de

capacité des infrastructures aériennes et au caractère de long terme de l'investissement visant à accroître cette capacité, il recommande, d'avoir recours à une tarification adaptée à une situation de congestion. À travers cette méthode, chacun prend en compte les coûts externes des retards.

L'apport essentiel de l'auteur avec cet article repose sur la mise en évidence d'une grande différence entre les aéroports et la route. Alors qu'un usager de la route représente une part infime du trafic sur la route, une compagnie aérienne représente une large part du trafic dans un aéroport. L'auteur prend l'exemple des aéroports congestionnés aux États-Unis qui sont exploités par une ou deux compagnies. Or, quand une compagnie décide d'exploiter un avion supplémentaire, elle prend en compte la congestion qu'elle s'impose à elle-même avec cet accroissement du trafic. L'idée de l'article est donc de développer un modèle de congestion en considérant que les usagers représentent une large part du marché.

L'analyse de la congestion aéroportuaire de BRUECKNER (2002) repose sur plusieurs hypothèses. D'abord, il est fait abstraction de toute considération « réseau » : un seul aéroport congestionné est étudié, les autres aéroports avec lesquels les compagnies opèrent leurs liaisons sont supposés avoir une capacité suffisante. Ensuite, deux périodes sont distinguées : une de pointe et une de hors pointe. Enfin, les retards imposent aux passagers des coûts liés au temps.

Outre la différence déjà mentionnée entre l'air et la route, d'autres jouent un rôle dans le modèle. Une différence économique concerne le coût de circulation. Celui d'une voiture est indépendant du niveau de la congestion routière, alors que celui d'un vol est largement affecté par la congestion aérienne. La congestion aéroportuaire réduit le temps journalier d'utilisation des avions, ce qui constitue un coût d'opportunité important. De plus, le personnel est mobilisé plus longtemps, ce qui accroît les coûts du travail, et lorsque les retards prolongent le temps passé dans les airs, les coûts du carburant augmentent également. Une autre différence vient du fait que les individus sur route conduisent leur propre véhicule, alors que les passagers aériens achètent un service de transport auprès d'une compagnie, pour lequel ils acquittent le prix d'un billet.

Étant donné la première différence citée entre les aéroports et la route, l'analyse de la congestion aéroportuaire aboutit à des résultats différents selon la structure du marché.

Commençons par étudier le cas du monopole, où une seule compagnie a un pouvoir de marché total sur l'aéroport. Elle fixe seule les prix de ses billets de pointe et de hors pointe. Les passagers, en comparant le bénéfice brut de leur voyage auquel ils retranchent le coût lié au temps et le prix du billet, pour les deux périodes, choisissent le moment auquel ils préfèrent voyager. Le prix du billet hors pointe est déterminé de manière à ce que le passager avec le plus faible surplus net soit indifférent entre voyager et ne pas voyager. Le prix du billet de pointe est quant à lui fixé selon la manière dont la compagnie veut allouer le trafic entre les deux périodes.

Le nombre de passagers en pointe à l'équilibre est celui qui maximise le profit de la compagnie. Le profit le plus élevé possible est obtenu lorsque les gains et les pertes dus au passage d'un avion de la période de hors pointe à celle de pointe se compensent. Or, en faisant passer des vols de la période de hors pointe à celle de pointe, trois effets sont attendus. Le

premier effet est une hausse des recettes, car plus de passagers payent plus cher. Un second effet est une baisse des recettes, car le prix du billet de pointe doit être diminué pour attirer des passagers et pour compenser leurs coûts liés au temps. Un troisième effet est une hausse des coûts opérationnels de la compagnie, due à la congestion.

L'équilibre sera donc atteint à condition que la recette supplémentaire générée par le premier effet égalise la baisse de recettes due aux deuxième et troisième effets. Étant donné que la compagnie est seule sur le marché, elle internalise tous les effets de la congestion. Son niveau de trafic est donc socialement optimal. Dans ce cas monopolistique, BRUECKNER (2002) considère qu'une « surtaxe » liée à la pointe de l'aéroport est inutile : elle risquerait de diminuer le niveau de trafic optimal pendant la pointe.

À l'opposé de cette situation se trouve la structure de marché parfaitement concurrentielle. Les prix des billets ne sont alors plus sous le contrôle des compagnies ; elles sont appelées « *price-taker* ».

Dans ce cas, les compagnies ne considèrent pas la hausse des coûts opérationnels qu'elles engendrent chez leurs concurrents lorsqu'elles programment un vol en période de pointe. De plus, elles ne prennent pas en compte les coûts supportés par les passagers. Étant donné la présence d'une multitude de compagnies sur le marché, l'ensemble des coûts de la congestion est donc ignoré par les compagnies lorsqu'elles prennent des décisions aboutissant à un accroissement de trafic.

Ce cas rejoint celui de la congestion routière. Sur un marché concurrentiel, BRUECKNER (2002) préconise donc des « surtaxes » qui reflètent l'ensemble des coûts externes engendrés.

Entre ces deux situations extrêmes, il existe le cas intermédiaire de l'oligopole. Seulement quelques compagnies sont présentes sur la marché et chacune détient un pouvoir de marché non négligeable.

À nouveau, lorsqu'une compagnie veut faire passer des passagers de la période de hors pointe à celle de pointe, les conséquences de ce déplacement sont triples. Par rapport au cas monopolistique, nous retrouvons les trois même effets. Cependant, bien que la baisse des prix des billets pour la période de pointe soit proposée à l'ensemble des passagers, la baisse des recettes qui en résulte sous les deuxième et troisième effets ne tient pas compte de tous les passagers, ni de toutes les compagnies. La hausse des coûts liés au temps pour les passagers des autres compagnies est ignorée, comme la hausse des coûts opérationnels des autres compagnies. Ainsi, seulement une partie des coûts externes est internalisée.

Dans le cas oligopolistique, BRUECKNER (2002) recommande donc une « surtaxe » d'une ampleur moins forte que dans le cas de la concurrence, le péage optimal ne portant que sur la partie non internalisée.

Par comparaison avec la route, une « surtaxe » liée à la congestion pour l'accès aux aéroports, ne doit pas faire payer tous les coûts externes, mais seulement ceux imposés aux autres. Cette surtaxe doit être inversement proportionnel au pouvoir de marché de la compagnie.

2.2.3 Justification de la tarification de pointe par son pouvoir incitatif

Le rationnement par les prix consiste à faire dépendre les prix des disponibilités à payer des agents. La dispersion des disponibilités à payer justifie un étalement des prix. Ce rationnement peut être mis en place de deux manières : soit selon l'ordre croissant des disponibilités à payer, soit selon leur ordre décroissant. Avec une tarification de pointe fondée sur les coûts de congestion, le rationnement par les prix se fait selon l'ordre croissant des disponibilités à payer. Celui qui est servi dans les temps paye en plus du coût marginal opérationnel, le coût marginal de la congestion, c'est-à-dire le coût qu'il impose aux autres agents du fait de sa demande. Ce coût correspond à la disponibilité à payer des agents non servis à temps. L'agent dont la demande est satisfaite « indemnise » la société en payant ce que les autres étaient prêts à payer pour être servis dans les temps. Ce sont les agents imposant les coûts de congestion les plus élevés, en privant les utilisateurs à forte disponibilité à payer de l'infrastructure, qui sont rationnés. Cette situation est celle de la section précédente (2.2.2).

Nous allons maintenant étudier un rationnement de la demande dans l'ordre décroissant des disponibilités à payer. Dans ce cas, les agents rationnés sont ceux pour lesquels les prix, toujours déterminés en fonction des disponibilités à payer mais d'une manière inverse par rapport au cas précédent, sont plus élevés. Désormais le prix pour un agent s'établit en fonction de sa propre disponibilité à payer. Les prix s'appuient donc sur la discrimination tarifaire (cf. section 1.2).

Avant de présenter une forme particulière de tarification de pointe, le *priority pricing*, nous discuterons du caractère discriminatoire des prix de pointe.

2.2.3.1 Le caractère discriminatoire d'une tarification de pointe

La tarification de pointe, qui ne repose pas sur un écart de coûts, peut être considérée comme une discrimination au second ou au troisième degré. Deux objectifs peuvent être poursuivis dans le cadre d'une tarification de pointe discriminante. Le premier peut être d'augmenter le prix pour une catégorie d'utilisateurs, afin de baisser le prix pour d'autres agents, qui avec un prix unique n'auraient pas pu bénéficier de ce service. L'infrastructure est rendue ainsi plus accessible. Un autre but peut être de vouloir écrêter les pointes.

Une manière d'atteindre ce dernier objectif est la discrimination au second degré. L'opérateur attend des utilisateurs qu'ils s'auto-sélectionnent parmi plusieurs couples « prix-période d'utilisation » proposés.

La discrimination au troisième degré peut quant à elle permettre d'atteindre les deux objectifs cités. Selon l'élasticité-prix des demandes des utilisateurs visés, l'un ou l'autre des buts sera atteint. Si cette élasticité est très forte, c'est-à-dire que la demande réagit beaucoup à des changements de prix, une tarification reposant sur une discrimination au troisième degré lissera les pointes. Elle découragera les utilisateurs d'exprimer leur demande en période de pointe. À l'inverse si l'élasticité-prix de la demande est faible, une tarification de pointe ciblée sur des caractéristiques exogènes des agents, n'aura que peu d'incidence sur la demande. Dans ce cas, les recettes supplémentaires retirées des prix de pointe pourraient être utilisées au profit d'autres utilisateurs.

Si l'on cherche à encourager la demande à se déplacer des périodes de pointe vers les périodes creuses, une discrimination au second degré qui laissera le choix aux agents ou une discrimination au troisième degré reposant sur des caractéristiques exogènes des agents peuvent être envisagées.

Avec une discrimination liée au choix des agents, en fixant un prix égal à la k^{e} disponibilité à payer la plus forte pour une capacité de l'infrastructure égale à K, l'opérateur incite les utilisateurs prêts à payer moins cher à exprimer leur demande, à une période moins chargée. La difficulté de cette solution théorique réside dans son application pratique. La k^{e} disponibilité à payer la plus forte étant ignoré *a priori* peut être obtenue par tâtonnement, en essayant plusieurs niveaux de prix.

Pour une discrimination liée à un signal exogène, l'opérateur doit identifier les caractéristiques communes aux utilisateurs qui présentent une demande individuelle similaire.

2.2.3.2 La tarification à la priorité

Une autre manière de rationner la demande par les prix, selon l'ordre décroissant des disponibilités à payer, peut être de proposer plusieurs prix associés à différentes priorités. Les agents dont la disponibilité à payer est forte obtiennent alors le service en priorité. Cette discrimination tarifaire s'appuie sur un aspect de la qualité : l'attente. Plus il faut attendre, moins le bien est cher. Cette pratique n'a d'intérêt que lorsque le producteur dispose d'une offre limitée et se trouve confronté à une demande qui dépasse la capacité.

Le *priority pricing* est un système de prix méconnu. Pourtant, il est pratiqué dans des secteurs dont l'utilité est quasi-quotidienne. Le *priority pricing* s'apparente à une différenciation verticale. La différence porte sur la qualité du service, exprimée en termes de temps d'attente pour être servi. Il s'avère qu'un système de *priority pricing*, qui combine prix-allocation, est plus efficace qu'un rationnement aléatoire avec prix uniforme.

Le *priority pricing* s'applique dans beaucoup de secteurs. Il consiste à faire payer un bien ou un service en fonction du temps passé avant de l'obtenir. L'existence de plusieurs tarifs dans l'acheminement du courrier par la Poste en est un exemple. Un expéditeur qui choisit une lettre avec un service « lent » accepte que la durée de distribution de sa lettre soit plus longue que celle avec un service « rapide », en contrepartie d'un prix plus faible. Il en va de même pour les soldes. Les prix diminuent au rythme des démarques. Les agents qui paieront des prix bas auront reporté leur achat dans le temps, leur probabilité d'obtenir le bien ayant alors diminué. Un autre exemple est celui de la fourniture de gaz ou d'électricité. Aux États-Unis, les compagnies d'électricité proposent des prix bas, et les agents qui les acceptent, subiront en premier des coupures dans l'approvisionnement en cas d'insuffisance de l'offre.

Les trois exemples précédents illustrent bien les différentes catégories d'« objets » pour lesquelles le *priority pricing* peut s'appliquer. Il s'agit de biens, auxquels les agents qui en obtiennent une unité attribuent une valeur, et de services, auxquels la valeur accordée est soit proportionnelle au temps d'utilisation, soit associée à une tâche spécifique. Le *priority pricing* intervient alors dans deux contextes différents. L'un d'eux est tel que le temps nécessaire pour satisfaire un consommateur est ignoré. Le bien ou le service vendu est considéré comme un flux : à chaque instant, chaque unité obtenue par un agent est immédiatement servie. C'est

le cas des biens vendus en solde et des services, tels que l'électricité pour le chauffage ou le téléphone, dont la valeur sur une période de temps est proportionnelle à la durée de la consommation. Dans l'autre contexte, le temps nécessaire pour satisfaire un consommateur est pris en compte. Il concerne les services qui nécessitent l'exécution d'une tâche spécifique dont la durée est fixe. Le service postal pour l'acheminement d'une lettre en est un exemple, comme l'électricité pour l'accomplissement d'une tâche manufacturière ou les télécommunications pour la transmission d'un message.

Dans le premier cas, celui d'un flux, le principe d'un système de priorités consiste à proposer à des clients différents prix, tels que ceux qui obtiennent le bien ou le service aux prix les plus élevés aient une probabilité d'être servis plus forte que les autres disposés à payer moins cher. Dans l'autre cas, celui d'un stock, les priorités se manifestent différemment. Les usagers qui paient plus cher bénéficient du service en priorité, c'est-à-dire avant ceux dont le prix acquitté est plus faible. Ces derniers pourront profiter du service, mais plus tard, ils subissent des retards plus ou moins longs, en fonction de leur rang de priorité. Un tel mécanisme de prix se modélise avec un système de file d'attente.

Les prix sont donc un outil qui sert à allouer la capacité du service et à déterminer les priorités. Le *priority pricing* correspond à une tarification incitative qui va conduire les clients à se comporter de telle sorte que leurs objectifs individuels vont coïncider avec ceux du système dans sa totalité.

Ce système de prix peut être assimilé à une forme de tarification de pointe. Par définition, les moments auxquels le *priority pricing* est pertinent coïncident avec les périodes de pointe, c'est-à-dire lorsque la demande se heurte à la contrainte de capacité. La tarification à la priorité a l'avantage d'être plus flexible. Elle autorise des agents à plus faible disponibilité à payer à être servis en période de pointe lorsque finalement la demande exprimée à cette période n'est pas très élevée. Cependant, les agents qui souhaitent connaître de façon certaine la période à laquelle ils seront servis préfèreront une tarification de pointe qui les incite à exprimer leur demande en période creuse plutôt qu'une tarification à la priorité qui ne peut leur permettre d'être servis en période de pointe qu'avec une faible probabilité.

Conclusion

À travers cette présentation théorique de la tarification de pointe, nous venons de voir que le choix de la modalité des prix se fait en fonction du caractère déterministe ou aléatoire de la demande.

Lorsque la demande est déterministe, il est possible de toujours la satisfaire, même si elle est très élevée. Les prix peuvent alors reposer sur les coûts de l'opérateur. L'intérêt d'une tarification de pointe, dans ce cas, est de faire payer un prix plus élevé aux utilisateurs de l'infrastructure pour lesquels le sur-dimensionnement a été mis en place. La justification de ce système de prix reposerait sur la base du principe du « coût du service » : l'utilisateur devant payer les coûts qu'il engendre. Mais, même si les coûts et les prix sont liés, de la discrimination tarifaire peut se justifier, avec des demandes de pointe et de hors pointe très proches. La demande des périodes creuses doit alors contribuer à financer la capacité des

périodes de pointe.

Dans l'autre situation, celle d'une demande aléatoire, toute la demande ne peut être satisfaite, elle doit être rationnée. Mis à part un rationnement aléatoire, le rationnement passe par une tarification de pointe reposant sur la base du principe de la « valeur du service ». L'intérêt, ici, est de faire prendre conscience aux utilisateurs du sur-coût qu'ils font supporter aux autres, en faisant en sorte que les externalités soient intégrées dans les calculs individuels. Un autre intérêt est de faire payer aux utilisateurs la valeur qu'ils accordent à l'infrastructure. Seul le choix entre un rationnement selon l'ordre croissant des disponibilités à payer et selon leur ordre décroissant est discrétionnaire, lorsque la demande est aléatoire.

La figure (2.5) résume toutes ces configurations.

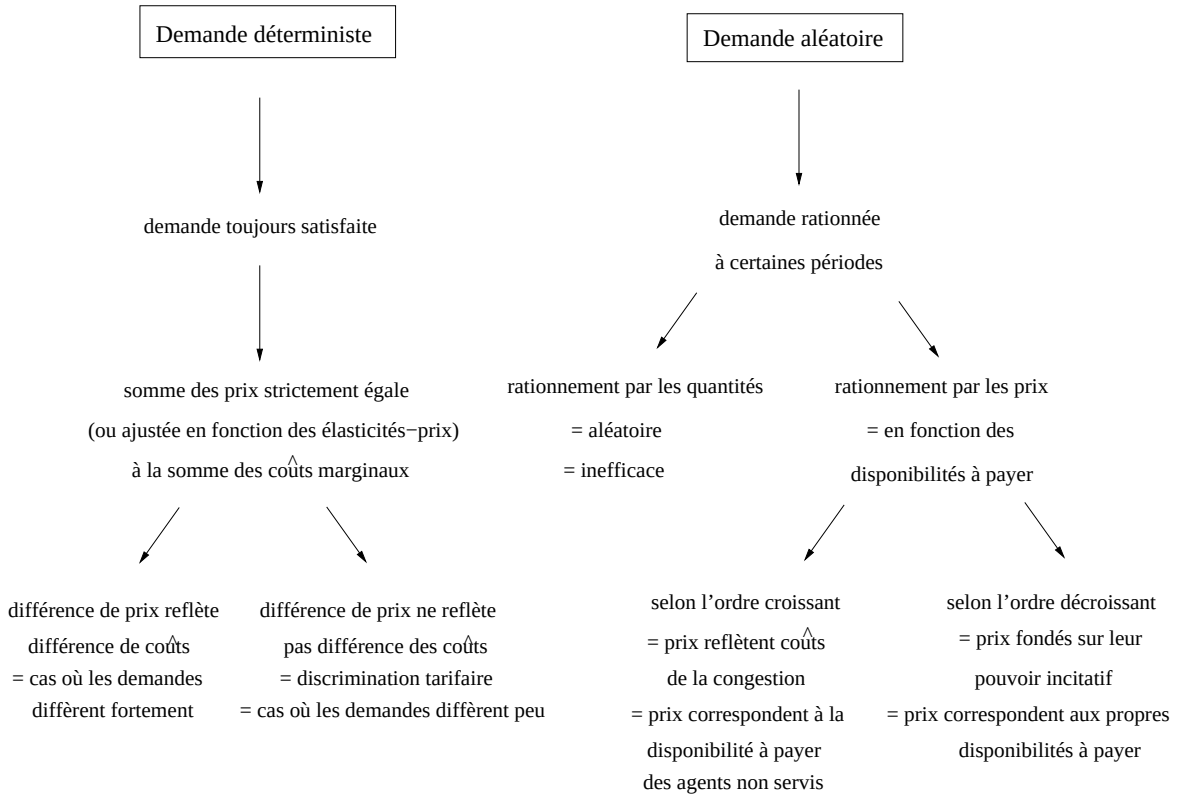

FIG. 2.5 – *Résumé de la tarification de pointe avec une offre fixe et une demande fonction du temps.*

Ainsi, une tarification de pointe n'aboutit quasiment jamais à des prix où chacun paie une contribution en proportion de la capacité, des coûts physiques, qu'il impose. Soit même les utilisateurs de la période creuse payent une partie des coûts, soit les coûts de congestion sont inclus dans les coûts, soit il faut discriminer entre les agents pour les inciter à déplacer leur demande.

Chapitre 3

Expériences de tarification de pointe

Introduction

De nombreux aéroports pratiquent la modulation tarifaire horaire. Les principales tranches horaires sur-tarifées sont la nuit et les heures en dehors des périodes normales d'activité de l'aéroport. À travers ces périodes, on comprend aisément que ces modulations tarifaires n'ont rien de commun avec la tarification de pointe. Elles se justifient par des coûts plus importants, soit des coûts directs, notamment la nuit en raison de l'éclairage, soit des coûts indirects, comme les nuisances sonores.

Les modulations tarifaires liées aux pointes sont en revanche beaucoup moins pratiquées.

En matière de redevances aéroportuaires, comme dans beaucoup d'autres domaines, les anglo-saxons ont été les précurseurs d'une modulation horaire. Les aéroports new-yorkais en 1968, puis les aéroports londoniens en 1972 ont décidé de faire dépendre leurs redevances de l'heure de l'atterrissage ou du décollage, en raison de problèmes de congestion.

Les deux chapitres précédents nous ont présenté différentes théories en matière de tarification d'infrastructure de pointe. Ces théories soulèvent de nombreuses interrogations sur leurs mises en application.

Les expériences new-yorkaises, londoniennes et quelques autres peuvent nous éclairer sur la manière de surmonter certaines difficultés.

Non seulement les expériences étrangères en matière de redevances aéroportuaires peuvent nous fournir des enseignements, mais aussi les expériences dans d'autres domaines d'activités.

Nous verrons dans un premier temps comment la tarification de pointe s'applique dans d'autres secteurs. Beaucoup d'entreprises de réseaux ont adopté ce genre de systèmes de prix. Nous présenterons ensuite les redevances aéroportuaires de pointe qui peuvent apporter des éléments de réponse à certaines questions.

3.1 Dans d'autres secteurs

Des industries de réseau, autres que les aéroports, pratiquent la tarification de pointe. Certaines ont fait ce choix depuis déjà longtemps.

C'est le cas pour l'énergie, avec EDF, pour les télécommunications, avec France Télécom, pour le transport, par la route, avec les sociétés d'autoroutes, ou par le fer. La situation ferroviaire est semble-t-il celle qui se rapproche le plus de l'aérien : les services et l'infrastructure sont séparés, gérés par des entreprises distinctes. La principale différence vient de l'absence de concurrence actuelle dans les services de transport ferroviaire.

Nous présentons les choix de modulation tarifaire des industries dans ces quatre secteurs.

3.1.1 L'expérience d'Électricité De France

Les coûts d'EDF présentent une certaine progressivité. Une règle de prix contingente à la structure des coûts est alors plus aisée à mettre en application. EDF dispose de plusieurs techniques de production. Le plan d'utilisation optimale des équipements est tel que la base de la production est réalisée par les générateurs dont le coût marginal de fonctionnement est le plus faible (hydraulique, nucléaire), mais dont le coût fixe est très élevé. Puis, pour satisfaire la demande dans les périodes où elle est plus importante, il fait appel aux générateurs utilisant d'autres technologies par ordre croissant du coût marginal (gaz, charbon). La figure (3.1) présente cette structure des coûts.

EDF propose donc une tarification qui reflète les différences de coûts et le caractère aléatoire de leur recours. L'option tarifaire TEMPO est définie en temps réel. L'année est divisée en trois types de jours aléatoires. Chaque jour est lui-même découpé en deux périodes fixes : les « heures pleines » et les « heures creuses ». Le type du jour est choisi par EDF la veille pour le lendemain. TEMPO a donc six prix différents qui se justifient par le recours à différentes énergies selon le niveau de la demande.

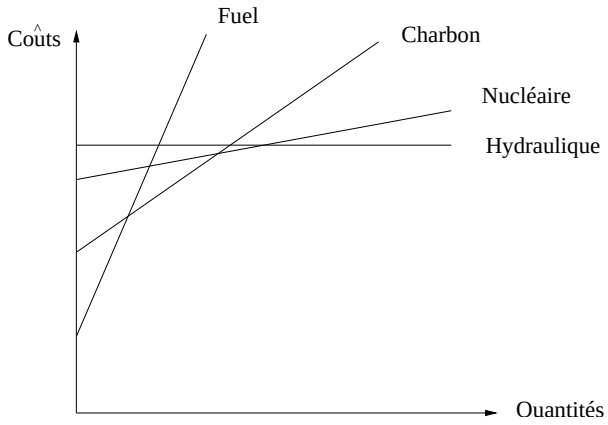

FIG. 3.1 – *Structure des coûts d'EDF*

3.1.2 L'expérience de France Télécom

En 1997, les services de téléphonie française ont été ouverts à la concurrence. France Télécom a continué à assurer ces services, mais a conservé son monopole lié à l'infrastructure. France Télécom fournit donc des services auprès des consommateurs finaux et assurent également un accès à son infrastructure pour d'autres opérateurs.

3.1.2.1 Les services de télécommunication

Comme pour EDF, il est possible de s'intéresser aux prix des services de télécommunication. France Télécom pratique des prix différents selon l'heure de la journée. Ainsi, les communications nationales des particuliers coûtent 50 % plus cher de 8 heures à 19 heures en semaine qu'aux autres moments. Pour les communications locales les heures pleines valent 83 % plus cher que les heures creuses.

Pour les professionnels, la période de pointe est élargie : elle s'étale de 7 heures à 22 heures. Pour les communications locales, le tarif de pointe est supérieur de 60 % au tarif hors pointe. En national, il est supérieur de 33 %.

3.1.2.2 L'infrastructure des télécommunications

L'exemple de France Télécom est également intéressant sur le plan de l'infrastructure. Les concurrents de l'opérateur historique et l'opérateur historique lui-même viennent s'interconnecter sur l'infrastructure : la partie locale du réseau de France Télécom. La question des charges d'interconnexion est importante.

Dans le cadre concurrentiel, la loi de réglementation des télécommunications impose à France Télécom de publier son catalogue d'interconnexion décrivant les principaux services d'interconnexion ainsi que les prix associés. Le gouvernement doit approuver cette offre et prend sa décision après examen de l'avis rendu par l'Autorité de Régulation des Télécommunications. L'ART indique que les tarifs d'interconnexion de France Télécom doivent être fondés sur les coûts.

France Télécom recouvre à travers ses tarifs d'interconnexion trois types de coûts : les coûts de réseau général, les coûts spécifiques et les coûts communs pertinents pour l'interconnexion. Les coûts de réseau général sont calculés sur la base de coûts incrémentaux de long terme. France Télécom les valorise sur la base de son modèle de comptabilité en coûts de remplacement. Les effets d'échelle et les effets du progrès technique ont été intégrés et se traduisent par une diminution des coûts unitaires. Les coûts spécifiques à l'interconnexion sont recouvrés dans leur intégralité sous forme d'un pourcentage uniforme appliqué aux coûts de réseau général (11 % pour 2003). La prise en compte des coûts communs pertinents s'est faite en 2003 par l'utilisation d'une majoration de 6,7 % appliquée à la fois aux coûts de réseau général et aux coûts spécifiques.

Les coûts de réseau général sont ensuite alloués par blocs de capacité. L'allocation retenue par France Télécom lui permet de recouvrer les coûts des différents éléments de réseau à un niveau compatible avec les coûts moyens incrémentaux de long terme. Les coûts spécifiques et les coûts communs sont alors alloués proportionnellement aux coûts de réseau général.

Les tarifs d'interconnexion comprennent une partie établie par appel et une autre établie par minute. Ces tarifs diffèrent selon l'heure de la journée. On distingue trois périodes :

- période bleue nuit : de 22 heures chaque jour à 7 heures du jour suivant ;
- période réduite : du lundi au vendredi de 7 à 8 heures et de 19 à 22 heures, ainsi que les samedis, dimanches et jours fériés de 7 à 22 heures ;
- période normale : le reste du temps.

Étant donné que le calcul des charges d'interconnexion repose sur les coûts, on en déduit que les coûts diffèrent selon ces périodes. Par rapport au tarif réduit, le tarif normal est supérieur de 55 % ; il est supérieur de 133 % au tarif bleu nuit.

3.1.3 Les expériences du réseau autoroutier français

Les sociétés chargées de l'exploitation des autoroutes en France ont elles aussi recours à la tarification de pointe.

3.1.3.1 SANEF

Sur l'autoroute A1 qui relie Lille et Paris, exploitée par la société SANEF, une modulation horaire est appliquée depuis le 26 avril 1992 pour les retours sur Paris le dimanche. Aux heures vertes (de 14 h 30 à 16 h 30 et de 20 h 30 à 22 h 30) les automobilistes bénéficient d'une réduction de 25 % par rapport au tarif normal du péage. En revanche, aux heures de tarif rouge (de 16 h 30 à 20 h 30), le péage est majoré de 25 %.

Grâce à cette modulation, 10 % des automobilistes qui circulaient aux heures de pointes ont décalé leur voyage vers d'autres heures. Ce déplacement de la demande a suffi à écrêter la pointe de trafic, qui se produisait traditionnellement à 19 heures. L'objectif d'étalement des retours sur Paris est donc atteint grâce à cette modulation.

3.1.3.2 Cofiroute

Cofiroute, responsable du réseau autoroutier du Sud-ouest, de l'Ouest et du Centre a également mis en place une modulation tarifaire, le dimanche pour les retours de week-ends aux abords de Paris. La modulation s'appliquait du dimanche 9 heures au lundi 3 heures, avec quatre tarifs qui s'échelonnaient de - 35 % à + 25 % du tarif standard.

Ce système a eu pour conséquence une réduction de trafic de 8 à 12 % en période de pointe. Les conditions de circulation se sont améliorées, la formation de « bouchons » étant nettement en diminution. Cependant, pour des raisons d'équilibre financier, l'opération mise en place du 24 mars au 25 novembre 1996, n'a pas été reconduite.

3.1.3.3 AREA

La société AREA des autoroutes de Rhône-Alpes a mené plusieurs hivers de suite, de 1993 à 1997, une opération qualifiée de « destination neige ». Le but de cette opération était d'inciter les habitants de la région à partir tôt le matin en direction des stations et revenir

tard le soir les samedis de vacances de sports d'hiver afin d'éviter des encombrements. Les automobilistes voyageant de 5 heures à 7 heures et de 19 heures à 21 heures profitaient d'une remise de cadeaux la première année et de tickets retours gratuits les années suivantes.

Ces remises, correspondant à des réductions de 50 %, eurent un grand succès.

3.1.4 Les expériences dans le ferroviaire français

Comme dans l'aérien, l'organisation du système ferroviaire français est séparée, avec d'une part, les services et d'autre part, l'infrastructure. Mais contrairement aux télécommunications, cette séparation, mise en œuvre afin de gérer la dette de la SNCF et en prévision d'une ouverture à la concurrence du transport ferroviaire, s'est opérée à travers un démantèlement vertical. En 1997, la SNCF a été divisée en deux entreprises : la SNCF, chargée les services, et Réseau Ferré de France (RFF) chargé de l'infrastructure. RFF est devenu un « fournisseur » de la SNCF, dans le sens où il lui fournit un input. RFF et la SNCF présentent tous les deux des tarifs soumis à une modulation horaire.

3.1.4.1 Les services de transport

Les prix pratiqués par la SNCF s'avèrent complexes. Alors que le tarif unique de la première classe garantit la clarté, les tarifs de seconde classe sont plus difficiles à déchiffrer. Les heures de circulation des trains introduisent une répartition en deux périodes. La période « normale » correspond aux heures creuses, et celle « de pointe » aux heures où il existe une forte demande. Cette période couvre généralement les vendredis et dimanches soirs, ainsi que les lundis matins, pour les départs et les retours de week-end. Cependant, pour certaines liaisons, la période de pointe peut être plus étendue. Elle varie notamment en fonction du sens de la liaison (en distinguant par exemple le sens Paris-Province de celui de Province-Paris) et peut s'appliquer tous les jours, par exemple tous les soirs de 17 heures à 19 heures.

L'optimisation de la recette, à travers le *yield management*, telle qu'elle est pratiquée par les compagnies aériennes est plus difficile à mettre en œuvre dans le transport ferroviaire. Contrairement aux avions, un trajet en train, présente plusieurs arrêts. Par conséquent, des passagers sont amenés à descendre ou à monter à bord du train. Ce va-et-vient de voyageurs s'intègre difficilement dans un modèle de gestion des prix et des capacités. Le résultat risque de ne pas être optimal.

Outre ces nombreuses modulations de prix, d'autres conditions tarifaires, rendent le système de la SNCF peu lisible ; certains le qualifient même « d'opaque ».

3.1.4.2 L'infrastructure ferroviaire

L'entreprise gestionnaire de l'infrastructure ferroviaire, RFF, a également adopté des redevances pour l'utilisation du réseau modulables selon les heures. Les barèmes des redevances sont calculés d'après trois critères :

– le droit d'accès : c'est en quelque sorte un abonnement, visant à couvrir les coûts engagés pour gérer le plan de transport (préparation des sillons et gestion de l'informa-

tion) ;
- le droit de réservation : il est fonction du niveau de trafic, de la consommation de capacité, des périodes horaires et de la qualité des sillons ;
- le droit de circulation : il est acquitté pour les circulations effectives et correspond approximativement à des coûts marginaux d'usage.

Les droits d'accès diffèrent selon la catégorie de la section considérée. Les sections sont réparties, selon le type de la ligne, en quatre catégories, elles-mêmes subdivisées en plusieurs sous-catégories. Les droits de circulation dépendent du type de transport : passagers ou marchandises. C'est pour le droit de réservation des sillons [1] que le prix unitaire par sillon-kilomètre varie selon l'heure. Une journée est divisée en trois périodes :

- heures creuses : de 0 h 30 à 4 h 30 ;
- heures normales : de 4 h 30 à 6 h 30, de 9 heures à 17 heures et de 20 heures à 0 h 30 ;
- heures pleines : de 6 h 30 à 9 heures et de 17 heures à 20 heures.

Pour les lignes périurbaines et les lignes à grande vitesse, le prix unitaire du droit de réservation des sillons est de 2 à 3 fois supérieur aux heures normales par rapport aux heures creuses, et également de 2 à 3 fois supérieur aux heures pleines par rapport aux heures normales. Pour les lignes interurbaines et les autres lignes, ce droit est nul ou quasi-nul avec un très faible niveau pour les heures normales et pleines.

L'un des objectifs de ces redevances est de favoriser la meilleure utilisation possible du réseau ferré.

3.2 Dans le domaine aéroportuaire

Les recommandations de la théorie en matière de redevances de pointe pour les aéroports suscitent de nombreuses interrogations. Comment instaurer un péage optimal, modifiable en permanence pour égaliser les coûts privés des utilisateurs ? Comment allouer correctement les coûts de la capacité ? Comment anticiper les effets d'une modulation tarifaire sur la demande ? Comment faire accepter des prix qui risquent d'exclure certains types d'aviation, ou de pénaliser les compagnies à plus faible pouvoir de marché ?

Nous recensons dans cette section les interrogations qui peuvent survenir lors du passage de la théorie à la pratique. Nous verrons quelles adaptations des résultats théoriques peuvent être faites et quelles sont les difficultés qui ne sont pas levées pour le moment.

3.2.1 Péage optimal et effet de seuil

Les modèles de congestion, dans lesquels les coûts sont internalisés définissent un péage optimal. Ce péage doit amener chaque utilisateur à avoir des coûts identiques. Certains ne supportent pas tous les coûts qu'ils imposent à autrui. Le péage sert à rééquilibrer la situation. Le péage optimal ainsi obtenu a alors la particularité de varier en continu, afin de s'adapter à

1. Un sillon correspond à la capacité d'infrastructure requise pour faire circuler un train donné d'un point à un autre à un moment donné. C'est l'équivalent d'un créneau dans le transport aérien.

chaque individu. Ceux qui imposent le plus de désagrément aux autres, mais qui supportent des coûts privés faibles car ils sont à l'heure se voient appliquer un péage élevé. À l'inverse, ceux qui sont confrontés à la congestion arrivent en retard et ne paient qu'un faible péage, voire aucun péage. Ainsi la tarification de pointe optimale définit un prix pour chaque instant, ce qui soumet les tarifs à une variabilité permanente.

Dans la pratique, il est très difficile de soumettre un tel péage aux utilisateurs. Une variabilité des prix forte et continue, risque de rendre le système peu lisible des utilisateurs et ainsi perdre ses effets incitatifs.

Parmi les aéroports londoniens qui ont adopté depuis 30 ans la modulation horaire, l'aéroport d'Heathrow se distingue des autres en raison d'une répartition du temps en trois périodes au lieu de deux généralement :

- période de « super pointe » : de 7 heures à 10 heures et de 17 heures à 19 heures d'avril à octobre,
- période de « pointe modérée » : de 5 heures à 7 heures d'avril à octobre,
- période de « hors-pointe » : le reste du temps.

Les redevances sont croissantes avec le degré de la pointe. La progression de ces redevances est intéressante, elle permet de se rapprocher d'un péage optimal.

Il existe un autre avantage à cet étalement des redevances. Il permet d'éviter des variations trop brutales de la période creuse à la période de pointe. Avec le palier intermédiaire des pointes modérées, les effets de seuil sont limités.

3.2.2 Information sur l'offre et la demande

Dans le cadre d'une tarification de pointe reposant sur la volonté de refléter la valeur accordée par les utilisateurs au service, la connaissance des coûts externes et de leurs élasticités-prix sont indispensables. Or, évaluer les conséquences financières de la congestion est une chose difficile à faire. Des informations sur la valeur du temps des passagers sont nécessaires.

Les compagnies aériennes, de leur côté, font face à des coûts directs et indirects. Souvent les compagnies doivent dédommager les passagers. Elles supportent également les coûts liés à leur organisation, notamment en *hub* qui nécessite de revoir la programmation des vols à la suite même d'un seul retard. Elles subissent aussi à plus long terme une perte de compétitivité, voire même des conditions de travail dégradées pour les employés. Ces coûts opérationnels sont des coûts directs. Les coûts indirects sont liés à l'anticipation des retards. En incorporant dans les programmes de vols des marges, afin de réduire l'impression de retard, et en disposant d'une flotte plus large, pour parer à des attentes trop longues dues aux interactions des vols, les compagnies augmentent leurs coûts opérationnels. L'évaluation de tous ces coûts est compliquée, et plus particulièrement celle des coûts indirects.

Lorsque l'on attend d'une tarification de pointe qu'elle incite une partie de la demande à se décaler vers une période moins chargée, il est également important de connaître l'élasticité-prix de la demande. Sans cette information, il n'est pas possible de mesurer l'ampleur de la

surcharge de pointe pour obtenir les effets souhaités.

Lorsque la tarification de pointe repose sur la volonté de refléter les coûts occasionnés par chaque utilisateur du service, le gestionnaire de l'infrastructure doit être en mesure d'identifier ces coûts. Or, dans son fonctionnement et son investissement, une infrastructure présente des coûts joints : il n'est pas possible de les attribuer directement à un service ou à un utilisateur. La difficulté est de les ventiler entre les différentes parties du réseau.

La BAA, qui a modifié ses redevances en 1972, justifie ce changement par le principe de la valeur du service et celui du coût du service. À partir de cette date, les taxes d'atterrissages ont inclus :

- un élément lié à la masse,
- un élément lié à la distance,
- un élément par passager, fonction de la distance, ne s'appliquant pas aux passagers en correspondance,
- une surcharge, fonction de la saison et du moment de la journée.

C'est l'élément lié à la distance qui reposait sur le principe de la valeur du service. La BAA considérait que plus un vol est long, plus la sensibilité de la compagnie aux redevances est faible. Les redevances pour les vols européens étaient multipliées par deux et celles pour les vols intercontinentaux l'étaient par un coefficient égal à quatre. Le système actuel ne tient plus compte de cet élément.

La surcharge, fonction de la saison et du moment de la journée, est fondée sur le principe du coût du service. Cette justification rejette alors l'idée de vouloir étaler la demande et s'appuie sur la volonté de mieux allouer les coûts. Pourtant, le contexte de congestion dans lequel ce nouveau mécanisme s'est mis en place laisserait plutôt penser que la BAA souhaitait internaliser les coûts liés à cette situation. Quoiqu'il en soit, nous ne savons pas quelle démarche pour le calcul des coûts a abouti aux taux pleins appliqués à partir de 1975. Nous ignorons quelle règle comptable a été utilisée.

3.2.3 Sensibilité aux prix

La faible sensibilité des compagnies aux variations des redevances s'expliquent de plusieurs manières. D'une part, lorsqu'une compagnie entreprend de desservir une région ou une ville, elle n'a pas réellement le choix de l'aéroport. Les alternatives sont peu nombreuses, voire inexistantes. Ainsi, une hausse des redevances ne modifiera pas nécessairement la demande des compagnies.

D'autre part, le revenu généré pour une compagnie lors d'un vol en période de pointe est très élevé. Elles-mêmes pratiquent la discrimination tarifaire en augmentant les prix des billets lorsque le demande est forte. L'écart de ce revenu par rapport à celui généré par un vol en période creuse est considérable. Ainsi, si on attend d'une modulation horaire des redevances qu'elle encourage une partie de la demande à se déplacer dans le temps, l'écart entre le prix standard et le prix de pointe doit être important.

Afin d'habituer les compagnies à la mise en place d'une redevance de pointe, son intro-

duction dans les aéroports londoniens fut progressive. Elle s'est étalée de 1972 à 1975.

Lors de l'instauration de prix de pointe, une surcharge de 20 £ a été appliquée de 8 heures à midi, d'abord du vendredi au lundi les mois de mai et juin, puis jusqu'en octobre tous les jours de la semaine. Une modification est ensuite intervenue pour la saison 1974. Elle a été marquée par l'introduction d'une surcharge encore plus importante, celle de la période de « *super-pointe* ». Les taux pleins, fixés pour l'année 1975 sont présentés sur la figure (3.2).

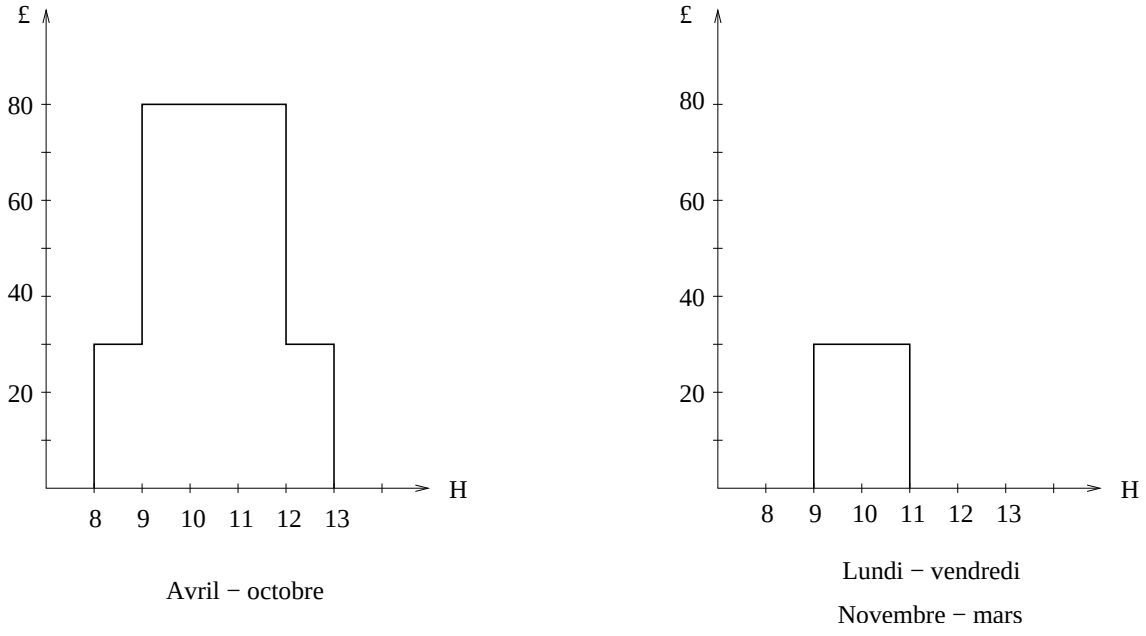

FIG. 3.2 – *Surtaxe de pointe.*

La surcharge de 80 £ a représenté un accroissement des taxes aéroportuaires de 87 % pour un vol domestique, de 55 % pour un vol européen, de 15 % pour un vol intercontinental avec un B 707 et de 7 % avec un B 747. L'effet était multiplié par deux dans les cas où l'atterrissage et le décollage se produisaient aux heures de pointe.

Du fait de cette surcharge, des recettes excessives par rapport aux coûts ont été levées. Elles ont servi à compenser une baisse de la taxe standard liée à la masse.

Depuis cette instauration définitive de redevances de pointe en 1975, les aéroports londoniens conduisent la même politique de prix. On observe que de 2002 à 2003, un réajustement a été opéré afin d'accroître l'impact de la surcharge. Les nouvelles redevances pour les 3 aéroports s'élèvent désormais aux sommes figurant dans le tableau (3.1).

3.2.4 Impact sur certains types d'opérations

Nous venons de voir que l'efficacité d'une tarification de pointe pour les aéroports dépend de l'amplitude des prix pratiqués. Le recours à la règle de prix qui reflète les coûts engendrés par chacune des demandes peut conduire à des prix de pointe très élevés. Outre le risque d'un déplacement de la demande, ce système peut poser un problème vis-à-vis de l'aviation

TAB. 3.1 – *Redevances d'atterrissage des aéroports londoniens en 2003.*

Aéroport		Périodes de pointe et nuits (£)				Périodes hors-pointe (£)			
	Avion à voilure fixe	Chapitre 2 surtaxé	Chapitre 3 majoré	Chapitre 3 de base	Chapitre 3 minoré	Chapitre 2 surtaxé	Chapitre 3 majoré	Chapitre 3 de base	Chapitre 3 minoré
Gatwick	de moins de 16 tonnes	385				95			
	de 16 à 55 tonnes	1 155	577,50	385	346,50	285	142,50	95	85,50
	de plus de 55 tonnes	1 155	577,50	385	346,5	375	187,50	125	112,50
Stansted	de moins de 16 tonnes	85				75			
	de 16 à 55 tonnes	390	195	130	117	285	142,50	95	85,50
	de 55 à 250 tonnes	630	315	210	189	360	180	120	108
	de plus de 250 tonnes	1 080	540	360	324	1 080	540	360	324
Heathrow	de moins de 16 tonnes	500				210			
	de 16 à 55 tonnes	1 500	750	500	450	630	315	210	189
	de plus de 55 tonnes	1 500	750	500	450	1 080	540	360	324
		Périodes de pointes modérées (£)							
	de moins de 16 tonnes	290							
	de 16 à 55 tonnes	870	435	290	261				
	de plus de 55 tonnes	1 455	727,50	485	436,50				

générale ou de tout autre type d'aviation dont les ressources financières sont limitées. La hausse des redevances aéronautiques va représenter une très forte augmentation des coûts pour certains types d'aviation. Adopter ce système reviendrait à en évincer du système. Pourtant, la théorie aboutit à ce résultat, car en période de congestion, la ressource, devenue rare, doit être allouée à ceux qui la valorisent le plus.

Certains aéroports appliquent une tarification de pointe qui touche indirectement l'aviation générale. D'autres ont choisi des modulations de tarif qui visent directement le transport de marchandises ou les opérations non commerciales.

3.2.4.1 Les expériences anglaises

Les premières années de la mise en œuvre de la tarification de pointe pour les aéroports londoniens ont permis d'observer son impact sur le trafic. Ce sont surtout les opérations non-commerciales ou de transport de marchandises qui furent affectées par la surcharge en raison de l'importance des coûts supplémentaires qu'elle a représentée.

À Manchester, les mesures visant à réduire la congestion, en période de pointe, ont pour cible avant tout les avions cargo. Les redevances d'atterrissage par tonne pour ces avions est de 3,27 £ et s'élève à 6,29 £ de 7 heures à 10 heures et de 16 heures à 19 heures.

Une mesure touche également les petits avions. Les avions transportant des personnes doivent payer une redevance par tonne (6,5 £). Cependant, de 7 heures à 11 heures du lundi au samedi et de 16 heures à 20 heures du lundi au vendredi, il existe une taxe minimale de 85 £. Ainsi, les avions pour lesquels le principe d'une redevance unique par tonne donne lieu à un montant inférieur à 85 £, c'est-à-dire les avions de faible masse, doivent s'acquitter d'une redevance plus élevée en période de pointe.

3.2.4.2 Les expériences américaines

L'aéroport de Boston a hésité à mettre en place une tarification de pointe. Il ne l'a pas fait face aux protestations de l'aviation générale.

À l'inverse, les aéroports new-yorkais confrontés à un sérieux problème de congestion (17 % des opérations retardées de plus de 30 minutes de 1960 à 1968) avaient pour cible ce type d'aviation. Le tableau (3.2) présente la part de l'aviation générale dans le trafic total à toutes heures et aux périodes de pointe en juillet 1968.

	Toutes heures	Heures de pointe
Kennedy	16,6	20,7
La Guardia	32,1	34,4
Newark	29,6	36,2
Total	25,0	30,0

TAB. 3.2 – *Part en % de l'aviation générale dans le trafic total en juillet 1968.*

Les trois aéroports new-yorkais ont souhaité agir sur l'aviation générale afin de réduire la congestion. Ils ont donc instauré à partir du 1er août une taxe pour les atterrissages et

les décollages entre 8 heures et 10 heures du lundi au vendredi et entre 15 heures et 20 heures chaque jour, pour les avions de moins de 25 places. Étant donné cette dernière restriction, cette surtaxe était clairement destinée aux vols de l'aviation générale. Les redevances s'élevaient dans les périodes de pointe à 25 $ au lieu de 5 $ en période normale. Les conséquences sur le trafic de l'aviation générale ont été immédiates, comme nous le montre le tableau (3.3).

	Toutes heures	Heures de pointe
Kennedy	- 7	- 18
La Guardia	- 28	- 41
Newark	- 17	- 30
Total	- 19	- 31

TAB. 3.3 – *Variation en % du trafic de l'aviation générale entre août et juillet 1968.*

Habituellement entre juillet et août, le niveau de trafic stagne. Cette année-là, il aurait même dû augmenter, étant donné qu'une grève des contrôleurs s'était terminée dans le courant du mois de juillet. On peut donc considérer que la chute du trafic de l'aviation générale est entièrement due à la surtaxe. Ce type de trafic diminuant, les retards également furent réduits.

3.2.4.3 L'expérience canadienne

L'aéroport de Toronto au Canada vise directement l'aviation générale : il surtaxe les appareils légers, de moins de 19 000 kilos, aux périodes de pointe. Les redevances d'atterrissage entre 7 heures et 10 heures et entre 14 h 30 et 21 heures du lundi au vendredi, et entre 14 h 30 et 21 heures le samedi, atteignent presque le double de celles aux autres heures : 140 dollars canadiens au lieu de 80.

Beaucoup d'autres aéroports instaurent une taxe minimale aux périodes les plus chargées afin de décourager les petits avions d'exprimer une demande à ces moments-là.

Cette discrimination selon le type d'appareil ou selon la fonctionnalité du vol risque de poser des problèmes d'acceptabilité au sein de certains aéroports.

3.2.5 Effets sur la concurrence

La proposition de BRUECKNER qui consiste à tenir compte du pouvoir de marché d'une compagnie dans un aéroport soulève également une difficulté. Selon sa démonstration les redevances devraient être corrélées négativement avec le pouvoir de marché, en raison de la prise en compte par une compagnie des effets néfastes de la congestion qu'elle s'impose à elle-même : plus une compagnie est implantée dans un aéroport, plus elle internalise les conséquences de sa décision, moins elle devra payer de redevances. Ce résultat pose un problème d'acceptabilité. On envisage mal qu'une mesure avantageant les compagnies à fort pouvoir de marché soit comprise de tous. Cette surtaxe risque d'être perçue comme un soutien à ces compagnies.

Cependant, la pratique dans les trois principaux aéroports de la région de New-York, est proche des recommandations de BRUECKNER. Les redevances sont fixées par unité de poids, avec des minima, fonction de la période et du caractère régulier du vol. Ainsi, les redevances minimales sont de 25 $ en période creuse et de 100 $ en période de pointe. Les compagnies aériennes et les vols qui relient quotidiennement l'aéroport voient ces charges minimum réduites. Elles sont de 20 $ et 50 $ respectivement en heure creuse et heure de pointe. Les opérateurs réguliers d'avion de petites tailles profitent donc de réductions. On conçoit aisément que ce sont ce type d'opérateurs qui détiennent un pouvoir de marché plus élevé que les autres sur l'aéroport.

Conclusion

À travers les expériences de redevances aéroportuaires de pointe présentées ici, beaucoup d'obstacles semblent levés. L'introduction progressive de prix de pointe paraît être de mise. Les problèmes d'acceptabilité sont surmontés car les choix qui pourraient être contestés sont justifiés par un accroissement du bien-être collectif. Il s'agit de les annoncer clairement et de manière lisible.

Cependant, un problème demeure, c'est celui de l'information. À notre époque, la détention d'information est source de richesse. À l'inverse, l'absence d'information est difficilement surmontable, car l'acquisition est par conséquent coûteuse. Des moyens doivent être mis en place pour obtenir une information pertinente et à moindre coûts.

Les effets d'une tarification de pointe sur la demande semblent positifs. Nous disposons ici surtout des conséquences les premiers temps de leur mise ne place. Mais l'ancienneté de tels systèmes pour les aéroports londoniens et ceux du New-York plaide en faveur de leur efficacité. On remarque, cependant, que les variations de prix entre périodes creuses et périodes de pointe doivent être sensibles pour provoquer des changements de demande.

Il est envisageable de prendre modèle sur certains secteurs. Mais tous ne sont pas comparables au système aéroportuaire. Par exemple, la structure progressive des coûts d'EDF ne correspond absolument pas à celle d'un aéroport. En revanche, dans une certaine mesure, une analogie peut être faite entre les redevances aéroportuaires et de navigation aérienne d'approche, et celles du réseau ferré.

Chapitre 4

La situation française

Introduction

L'objet de notre étude est la congestion aéroportuaire : nous cherchons à résorber ce problème, afin que l'offre aéroportuaire soit de meilleure qualité et ne pénalise pas la demande. Notre intérêt se porte donc sur les infrastructures aéroportuaires et les prestations qui y sont afférentes. Ces infrastructures sont essentiellement celles des installations au sol : pistes, voies de circulation, aires de stationnement, aérogares. Mais elles concernent aussi les infrastructures de navigation aérienne et les prestations de contrôle du trafic aérien.

Tous les services apportés aux compagnies aériennes et aux passagers dans les aéroports ou aux abords donnent lieu à la perception de redevances. Selon la prestation, ces redevances sont perçues par le gestionnaire de l'aéroport ou par la Direction de la Navigation Aérienne.

De nombreux éléments entrent en jeu dans la définition de ces redevances. Mais la pertinence de certains d'entre eux peut être discuté. Le prix du contrôle aérien est déterminé en fonction des coûts : tous les frais engagés sont entièrement couverts. L'inconvénient de cette règle est qu'elle n'encourage pas à maîtriser les coûts.

La masse de l'avion est un élément qui influence les redevances. Ce choix a été fait dans le but d'introduire une certaine « équité », en faisant payer les avions selon leur disponibilité à payer, avec l'hypothèse sous-jacente que la masse de l'avion est liée positivement à la propension à payer. Cette relation ne constitue pas une « bonne » incitation du point de vue de la congestion. Afin de réduire le trafic, il semblerait plus pertinent d'inciter les compagnies à travers les redevances à utiliser des avions plus gros, moins fréquemment.

Certains effets externes sont internalisés : ceux des nuisances sonores. Plus le niveau acoustique d'un avion est élevé, plus ces redevances sont fortes, l'effet étant accentué la nuit.

Cependant, une autre externalité négative demeure exclue du système des redevances aéroportuaires : c'est la congestion. Ce comportement à l'égard de cette situation est source d'inefficacités. La répartition des ressources aéroportuaires n'est pas optimale.

À partir des éléments théoriques développés dans les deux premiers chapitres et des enseignements tirés du troisième, nous devrions être en mesure d'envisager une tarification de

pointe pour les aéroports de Paris. Cependant, à ce stade, nous manquons d'informations cruciales pour aller plus en avant. Les recommandations en matière de prise en compte de la congestion dans les redevances aéroportuaires ne peuvent conduire à une application directe. Un travail sur les données est indispensable au préalable.

La première partie de ce chapitre est une analyse des redevances aéroportuaires actuelles. Les éléments négatifs de ces redevances au regard de la congestion seront soulignés. La deuxième partie met en avant les pistes à suivre désormais pour introduire une modulation horaire liée à la pointe dans les redevances des aéroports congestionnés en France.

4.1 Analyse des redevances aéronautiques françaises

Les redevances aéronautiques françaises répondent à certaines exigences ou recommandations des différentes autorités publiques en charge de prestations aériennes. La principale d'entre elles est la couverture des coûts. La référence aux coûts du gestionnaire de l'infrastructure est récurrente. Cependant, ce choix peut être discuté, notamment parce qu'il occulte l'existence d'autres coûts, non comptables, et qu'il néglige l'efficacité de l'allocation.

Après avoir présenté les différents types de redevances dues pour un atterrissage, nous adresserons quelques critiques à la définition de ces redevances. Des inefficacités liées à la congestion et par voie de conséquence au mode de rationnement apparaissent.

4.1.1 Modalités des redevances

Les principaux rôles joués par les aéroports auprès des avions concernent le contrôle aérien sur et près de l'aéroport, et la mise à disposition d'une infrastructure. Les contrôles d'approche et d'aérodrome, effectués par des contrôleurs à partir de la tour de contrôle, sont chargés respectivement de la circulation au sol sur la plate-forme, de la phase initiale de la montée et de la phase terminale de la descente des avions, pour l'un, et de la montée et de la descente des avions, pour l'autre. De son côté, l'infrastructure physique comprend essentiellement les pistes et une ou plusieurs aérogares.

Nous allons voir dans cette section, à quelles redevances tout ceci donne lieu. Il existe plusieurs redevances. Nous présenterons celles pour services terminaux, celles d'atterrissage, celles des passagers et celles de stationnement.

4.1.1.1 Redevance pour services terminaux

La Redevance pour Services Terminaux de la Circulation Aérienne (RSTCA) rémunère les services rendus au titre du contrôle d'approche et du contrôle d'aérodrome. Elle couvre les coûts occasionnés par les services de la circulation aérienne au départ et à l'arrivée. Le montant de la redevance due à chaque décollage se calcule de la façon suivante :

$$RSTCA_i = Tu \times 1,247 \times M_i^{0,9} \tag{4.1}$$

où M_i est la masse de l'avion i et Tu un taux unitaire.

Au 1er janvier 2003, ce taux unitaire était en France de 4,43 euros. Il est déterminé de manière à égaliser les coûts du contrôle d'approche et du contrôle d'aérodrome aux recettes.

Cette formule s'applique indifféremment à tous les aéroports soumis aux redevances pour services terminaux. On peut alors considérer que ces redevances sont discriminantes, car les coûts du contrôle d'approche et d'aérodrome ne sont pas identiques pour tous les aéroports. Il existe donc des subventions croisées entre aéroports.

L'introduction de la masse des avions est également un élément discriminant. Elle ne se justifie pas par une différence des coûts du contrôle selon la taille de l'avion. C'est plutôt le fruit d'une recommandation de l'OACI. Le conseil de l'aviation civile préconise de faire payer les avions en fonction de leur disponibilité à payer et considère qu'il existe « une relation approximative entre la capacité en charge payante et la masse de l'avion ».

4.1.1.2 Redevance d'atterrissage

La redevance d'atterrissage rémunère l'utilisation de la piste, ainsi que divers services. Comme les autres redevances aéroportuaires, mais à la différence des redevances pour services terminaux, la redevance d'atterrissage est perçue par le gestionnaire de l'infrastructure, généralement les chambres de commerce et d'industrie, sauf pour les aéroports de Paris pour lesquels elle est perçue par ADP.

Pour ces derniers, la redevance est fondée sur un tarif de base. Il est fonction de la masse. À ce tarif est ensuite appliqué un coefficient de modulation. Il dépend de différents éléments, dont le groupe acoustique de l'avion, l'heure et le lieu de l'atterrissage.

Il existe cinq niveaux acoustiques. Pour les trois premiers, les coefficients sont supérieurs à un. On cherche de cette manière à faire payer plus cher les avions dont les nuisances sonores sont élevées. Pour le niveau 4, le coefficient est égal à un et il est inférieur à un pour les avions de niveau 5. La modulation horaire distinguent deux périodes, celle du jour allant de 6 heures à 23 h 30 et celle de la nuit allant de 23 h 30 à 6 heures. La nuit, les redevances sont deux fois plus chères pour les avions des trois premiers niveaux. Pour le niveau 4, elles sont inchangées, et sont légèrement plus élevées pour les avions de niveau 5. Il existe des différences entre les coefficients de Roissy et d'Orly, seulement pour les trois premiers niveaux acoustiques, ceux d'Orly étant un peu plus élevés.

Ces redevances dépendent donc d'éléments qui influencent une partie des coûts. Les coûts des pistes et des aérogares sont fonction de la taille des avions qui y ont accès. La corrélation étant positive, on comprend que ces redevances varient avec la masse. Les coûts de l'éclairage la nuit peuvent aussi justifier des redevances plus élevées en période nocturne. Certains coûts externes sont aussi inclus dans le calcul. Plus les avions sont bruyants, plus ils doivent payer cher et l'effet sur les redevances est accentué la nuit, période où le bruit est encore plus nuisible. Les externalités négatives subies par les riverains des aéroports sont donc prises en compte.

4.1.1.3 Redevance des passagers

La redevance des passagers est due pour l'utilisation des aérogares et des installations aménagées sur les aéroports pour la réception des passagers. Elle est perçue à l'occasion de l'embarquement du passager.

À ADP, il s'agit d'un montant fixe par passager. Les prix varient selon la destination. Il existe quatre groupes de tarifs : pour les passagers à destination de la métropole, de l'Union Européenne zone Schengen, de l'Union Européenne hors zone Schengen et d'un aéroport international. Pour les passagers en correspondance, la redevance est diminuée de 10 %. L'argument avancé pour justifier cette réduction, est lié à une moindre utilisation de l'aérogare par ce type de passagers.

4.1.1.4 Redevance de stationnement

La redevance de stationnement d'ADP est fonction de plusieurs éléments :

- la masse de l'avion ;
- le type du poste de stationnement utilisé : aire de trafic au contact ou au large, ou aire de garage ;
- la durée d'occupation par l'avion de l'aire.

Seulement pour un stationnement sur une aire de trafic au contact, une part fixe par tonne, s'ajoute au montant de la part variable définie par tonne et par heure.

4.1.2 Critiques adressées aux redevances

La manière dont sont définies les redevances aéronautiques conduit à des inefficacités.

Une première critique adressée à ces redevances pourrait concerner les coûts. Les prix sont définis de façon à couvrir exactement les coûts sans qu'aucune contrainte ne pèse sur le gestionnaire de l'infrastructure. Cette absence de maîtrise des coûts ne fait cependant pas partie du cadre de notre étude. Nous insisterons donc essentiellement sur deux autres inefficacités.

La congestion aéroportuaire et le fait qu'elle ne soit pas prise en considération sont sources d'inefficacités. Le rationnement qui prévaut actuellement pour les aéroports congestionnés est lui aussi inefficace. La discussion de l'attribution de créneaux aéroportuaires est exclue de cette étude. Cependant, dans le cadre d'une tarification de pointe aéroportuaire, l'arbitrage entre un rationnement par les quantités et par les prix doit être évoqué.

4.1.2.1 Les inefficacités liées à la congestion

Aucun élément dans les redevances n'est introduit pour refléter la congestion. Pourtant il s'agit d'un problème récurrent dans certains aéroports. Une manière d'identifier les aéroports saturés peut être d'examiner leur taux de retard. Mais il existe une autre façon de connaître les aéroports dont la capacité est insuffisante : ceux qui font partie de la liste internationale des aéroports dits « coordonnés » sont des aéroports qui ne peuvent effectivement pas accueillir toute la demande qui se présente à eux.

La non internalisation des effets externes aboutit à une quantité sous-optimale. Lorsque les externalités sont positives, la quantité d'équilibre est insuffisante ; dans le cas opposé, elle est trop élevée. Ainsi, la congestion aéroportuaire, source d'effets externes négatifs, doit être prise en compte pour accroître le bien-être collectif.

4.1.2.2 Les inefficacités liées au rationnement par les quantités

Pour les aéroports congestionnés, le rationnement se fait par les quantités. Des créneaux sont attribués gratuitement selon « les droits du grand-père ». Lors de l'introduction de cette règle, les allocations passées sont prises comme référence. Par la suite, les créneaux libérés sont réalloués, en grande partie aux compagnies disposant déjà de créneaux et pour une autre partie à des compagnies souhaitant exploiter une nouvelle liaison.

Nous allons comparer le surplus social généré par ce rationnement avec celui d'un rationnement par les prix dit « parallèle ». Dans le second cas, le bien est alloué en priorité à ceux qui sont disposés à payer le plus pour l'obtenir.

On suppose qu'un service est disponible à un certain prix, que nous pouvons considérer nul sans perte de généralité. À ce prix, N agents sont intéressés par le service, mais seuls K agents, avec $K < N$, peuvent être servis en même temps. Chaque agent est caractérisé par une disponibilité à payer pour être dans les K premiers à bénéficier du service. L'agrégation de toutes ces valeurs conduit à une fonction de demande inverse, demande pour être servi en premier, qui exprime un prix décroissant en fonction d'une quantité d'agents servis :

$$P = D^{-1}(Q)$$

Si le rationnement est aléatoire, tous les agents ont la même probabilité de ne pas être servis immédiatement et de devoir attendre ; cette probabilité est égale à $\frac{N-K}{N}$. Le surplus social de cette situation est donc donné en espérance. Il est égal au produit de la probabilité d'être servi $\left(\frac{K}{N}\right)$ avec le surplus associé à une demande entièrement satisfaite :

$$\mathbf{E}(S_a) = \frac{K}{N} \int_0^N D^{-1}(q)dq \tag{4.2}$$

Dans le cas d'un rationnement parallèle, les premiers servis sont ceux qui sont les plus disposés à payer pour être servis en priorité. Le surplus social, associé à la situation dans laquelle les K agents ayant la plus forte disponibilité à payer peuvent bénéficier du service sans attendre, est :

$$S_p = \int_0^K D^{-1}(q)dq \tag{4.3}$$

Afin de savoir lequel de ces deux schémas de rationnement est le plus efficace, nous devons comparer les surplus des équations (4.2) et (4.3). Étant donné que la fonction de

demande inverse est décroissante, nous savons que :

$$\begin{aligned}&\int_0^K D^{-1}(q)dq + \int_K^{2K} D^{-1}(q)dq + \ldots + \int_{(\frac{N}{K}-1)K}^{N} D^{-1}(q)dq \\ &< \underbrace{\int_0^K D^{-1}(q)dq + \ldots + \int_0^K D^{-1}(q)dq}_{\frac{N}{K} \text{ fois}}\end{aligned} \tag{4.4}$$

chaque n^{e} terme du membre de gauche étant inférieur ou égal au n^{e} terme du membre de droite. L'écriture simplifiée de cette inégalité (4.4) est :

$$\int_0^N D^{-1}(q)dq < \frac{N}{K}\int_0^K D^{-1}(q)dq$$

elle-même équivalente à :

$$\frac{K}{N}\int_0^N D^{-1}(q)dq < \int_0^K D^{-1}(q)dq$$

soit

$$\mathbf{E}(S_a) < S_p$$

L'espérance du surplus social est donc plus importante avec un rationnement parallèle, qu'avec un rationnement proportionnel. C'est d'ailleurs pour cette raison que l'adjectif « efficace » est associé au schéma selon lequel les plus fortes disponibilités à payer sont servies en priorité.

La figure 4.1 est une illustration de la comparaison des surplus sociaux avec les deux règles de rationnement possibles. La fonction de demande inverse est représentée par une droite.

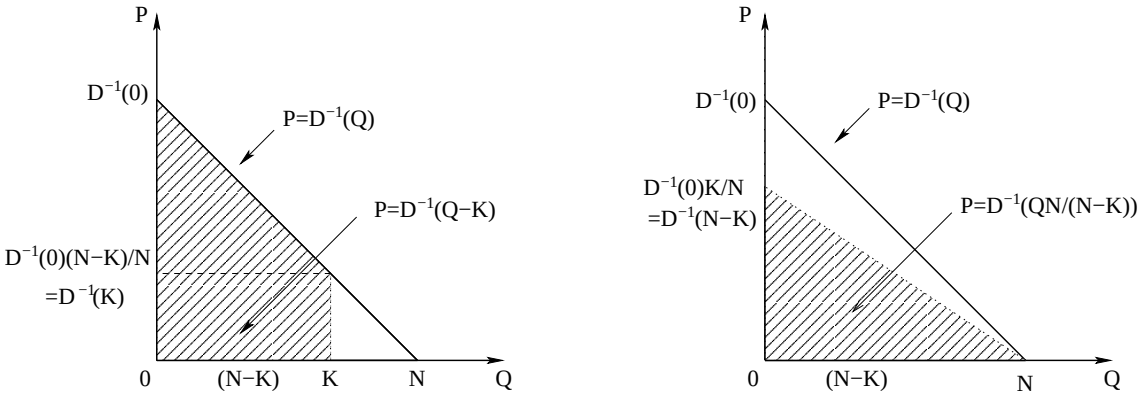

FIG. 4.1 – *Surplus social avec rationnements parallèle et aléatoire.*

Le premier cas est celui d'un rationnement parallèle. Les agents aux K plus fortes disponibilités à payer sont servis en premier. La demande des autres, $(N-K)$ agents, est reportée,

leur nouvelle demande est représentée par la droite la plus fine, « parallèle » à la fonction de demande inverse initiale. Le surplus social correspond à la partie hachurée.

L'autre cas est celui d'un rationnement proportionnel. La droite en pointillé indique l'espérance de la valeur de l'utilité retirée du service quand la probabilité de ne pas être servi est $\frac{N-K}{N}$. La partie hachurée sous cette droite est donc le surplus social obtenu avec un rationnement aléatoire.

D'après la démonstration précédente, la surface hachurée du premier cas est plus grande que celle du second cas.

4.2 Pistes pour l'avenir

Nous nous interrogeons désormais sur la justification à donner à une tarification de pointe pour les aéroports français coordonnés.

Par rapport aux éléments de théorie vus dans les chapitres précédents, nous devons identifier si la demande pour les infrastructures et prestations aéroportuaires est déterministe ou aléatoire. Ensuite, une fois cette analyse faite, nous devons approfondir notre information sur les coûts, qu'ils soient directs ou indirects. Dans le cas d'une demande aléatoire, une connaissance plus détaillée de la demande est indispensable.

4.2.1 Identification du caractère déterministe ou aléatoire de la demande

Les résultats théoriques de la tarification de pointe repose sur la distinction entre une demande à caractère déterministe et une demande à caractère aléatoire. Afin de savoir si une tarification de pointe se justifie sur la base des coûts de l'opérateur ou sur la base d'un rationnement efficace de la demande par les prix, la première étape est d'identifier de quel type est la demande. Cette identification nous permettra ensuite d'envisager la manière dont une tarification de pointe doit être mise en place.

4.2.1.1 Enseignement théorique

Si l'on considère que l'on connaît la demande de façon certaine, il est recommandé de satisfaire toute la demande. On augmente la capacité jusqu'à ce que le coût marginal opérationnel et celui de la hausse de la capacité rencontrent la courbe de demande. À ce point, la demande n'excédera pas l'offre.

Mais dans le cas où il existe un risque que la demande n'atteigne pas exactement le niveau anticipé, la stratégie doit être différente. Si la taille de l'infrastructure est trop grande, les coûts engagés ne seront pas recouvrés. Si au contraire la capacité est insuffisante, le système de prix reposant sur les coûts ne pourra pas remplir son rôle. Il est alors nécessaire de compléter ce mécanisme par un autre : le rationnement par les quantités ou par les prix.

4.2.1.2 État de la demande pour les infrastructures et prestations aéroportuaires

Il s'avère qu'il est difficile de connaître parfaitement la demande d'atterrissage. À moyen et long terme, des prévisions de trafic sont faites pour évaluer le niveau de la demande. Les redevances ne sont pas inclues dans ces modèles d'estimation comme variable explicative de la demande. Ces redevances couvrant strictement les coûts, le volume de trafic joue un rôle ; mais les compagnies sont considérées comme des « price-taker » : des agents qui ne sont pas susceptibles à eux-seuls d'influencer le niveau des prix. Le principal déterminant de la demande est l'activité économique. En période de forte croissance, la demande de transport aérien, et par conséquent la demande d'atterrissage est soutenue. À l'inverse, en période de ralentissement économique, le transport aérien connaît également un ralentissement de son activité.

À très court terme, ce sont divers incidents qui peuvent rendre la demande aléatoire. Ces modifications ne relèvent souvent même pas des compagnies, car il s'agit de retards qu'elles ne maîtrisent pas (conséquence d'autres retards ou retards de passagers) ou de problèmes techniques.

Il est donc difficile de disposer d'une connaissance exacte de la demande d'atterrissage pour toutes les périodes. On connaît surtout la tendance ; on sait quels sont les heures, les jours, les mois les plus demandés, mais on ignore le niveau précis de la demande.

Il semble donc raisonnable de considérer que la demande pour les prestations assurées dans le cadre des infrastructures aéroportuaires est aléatoire.

4.2.1.3 Recommandations

Nous avons déjà vu qu'une tarification de pointe n'est pas seulement utile lorsqu'un problème de capacité se pose, mais elle peut aussi servir quand il n'existe pas de contrainte sur la capacité : étant donné que les coûts fixes d'installation d'une infrastructure nécessitent une élévation du prix au-dessus du coût marginal, la tarification de pointe peut remplir efficacement ce rôle.

À partir du moment où la demande présente des pointes temporelles, on peut envisager de faire payer à une partie des utilisateurs l'accroissement de capacité qu'ils rendent nécessaire. Ainsi, même avec une demande aléatoire, une tarification de pointe se justifie, mais sur la base des coûts et non de la congestion.

Dans le cas où l'on souhaite traiter de la congestion à travers des redevances de pointe, ce qui est l'objet de cette étude, étant donné le caractère aléatoire de la demande, l'approche est différente. Avec une demande fluctuant dans le temps, on sait que la capacité ne sera pas suffisante en toute circonstance. La taille de l'infrastructure devra être celle qui rend égal le coût marginal au bénéfice marginal, l'excès de demande par rapport à la capacité n'étant pas servi dans les temps. Or, un rationnement efficace passe par les prix.

Dans ce cas, il faut envisager une discrimination tarifaire, fonction de la période à laquelle les prestations aéroportuaires sont assurées : la différence de prix ne se justifiera alors pas par une différence de coûts.

4.2.2 Étude approfondie des coûts

Nous avons vu que lorsqu'une tarification de pointe repose sur les coûts, ceux-ci peuvent être de natures différentes. On distingue les coûts du gestionnaire de l'aéroport des coûts de la congestion, supportés par les agents dont la demande n'est pas satisfaite dans les temps.

4.2.2.1 Les coûts de l'opérateur

Dans le cas où une tarification de pointe reflète des disparités dans les coûts engendrés par les utilisateurs de l'infrastructure auprès du gestionnaire, une étude approfondie des coûts est nécessaire, notamment l'identification de certains coûts présentés à la section 2.1.1.1.

Par exemple, pour identifier les coûts dus à la pointe, le calcul est le même que lorsqu'on évalue les coûts incrémentaux. Il faut calculer quel serait le niveau de la capacité en l'absence de la demande de pointe. L'écart de coût entre la capacité effective et cette capacité sera le sur-coût induit par la demande de pointe.

Dans l'exemple de Londres, la tarification de pointe est justifiée sur la base des coûts de l'opérateur. Il serait intéressant de disposer de la méthode utilisée par la BAA pour fonder sa tarification de pointe sur ses coûts. Une telle information risque cependant d'être difficile à obtenir.

Les États, comme les entreprises, divulguent très difficilement les données concernant leurs coûts. Même en France, dans le cadre d'une étude sur les coûts des opérateurs des infrastructures et des prestations aéroportuaires, le risque est fort d'être confronté à des difficultés dans le recueil des informations.

4.2.2.2 Les coûts de congestion

D'autres coûts, non pas supportés par le gestionnaire de l'aéroport, mais par les compagnies et les passagers qui subissent des retards du fait de l'infrastructure de capacité, doivent jouer un rôle dans la tarification de pointe. Ces coûts externes sont à l'origine d'inefficacités : ils diminuent le bien-être collectif. C'est à ce titre qu'ils doivent être pris en compte.

Cependant leur évaluation ne relève pas, comme pour les coûts de l'opérateur, d'un calcul comptable. Ces coûts de congestion sont supportés, comme nous l'avons vu dans la section 3.2.2, par les compagnies et leurs passagers. Ils se calculent à partir des valeurs accordées par ces différents agents au fait d'être servis dans les temps. Ces informations ne relèvent pas des coûts de l'opérateur, mais des agents eux-mêmes.

Ce sont donc des informations sur la demande dont nous avons besoin. Nous nous intéressons désormais aux fonctions de demande, mais nous ne discuterons que de celles des compagnies, clientes directes des aéroports.

4.2.3 Informations sur la demande

Pour élaborer de nouvelles redevances fonction des disponibilités à payer des compagnies, il est utile de disposer de nombreuses informations sur la demande. Mais justement,

ces informations sont difficiles à obtenir et il n'est pas dans l'intérêt des compagnies de les révéler.

4.2.3.1 L'utilité d'informations sur la demande

Une fois identifié le caractère aléatoire de la demande et reconnu l'intérêt d'instaurer une tarification de pointe visant à rationner la demande excessive par rapport à l'offre, deux méthodes de calcul des prix sont envisageables. Un choix doit être fait entre une tarification de pointe reposant sur les coûts de la congestion et une ayant un pouvoir incitatif. Ce choix est purement discrétionnaire, les deux méthodes aboutissant à un résultat optimal à condition de respecter les conditions présentées dans les modèles de congestion (cf. section 2.2.2) et celles liées à la discrimination tarifaire (cf. section 1.2).

Quel que soit le choix entre ces deux méthodes, il est nécessaire de disposer d'informations sur la demande. Il s'agit notamment de connaître la fonction de demande qui exprime la quantité demandée en fonction des prix.

La connaissance des disponibilités à payer des compagnies pour bénéficier des infrastructures et des prestations aéroportuaires dans les temps permet de mesurer les coûts de la congestion. Ces coûts sont égaux aux disponibilités à payer des compagnies non servies à l'heure souhaitée. Ils représentent la perte, appelée « coût d'opportunité », que cela représente pour eux.

La demande étant décroissante avec le prix, la connaissance du prix auquel la quantité demandée correspond à la capacité de l'infrastructure permet de fixer des redevances égales à ce prix et ainsi de décourager la demande excessive. Cette dernière est incitée à se déplacer vers une autre période moins chargée.

En disposant de plus d'informations sur la demande, l'autorité aéroportuaire serait en mesure d'anticiper les effets sur les différentes catégories de la demande. La crainte de pénaliser une catégorie de compagnies déjà fragiles ou la crainte de toucher une compagnie historique installée sur l'aéroport orientera le choix dans la manière de déterminer les redevances.

On a vu qu'aux États-Unis cet arbitrage avait été différent entre New-York et Boston, Boston voulant éviter de pénaliser l'aviation générale avec une tarification de pointe et New-York au contraire adoptant un mécanisme visant à réduire le trafic des petits avions en période de pointe.

L'orientation de la tarification de pointe relève donc d'un choix politique.

4.2.3.2 Le problème de révélation d'informations sur la demande

La principale difficulté de cette étape est la révélation d'informations. La fonction de demande est difficile à évaluer. Plusieurs manières de faire sont tout de même envisageables

Par exemple, pour une discrimination liée au choix des agents, la k^e plus forte disponibilité à payer peut être obtenue par tâtonnement. L'essai de plusieurs niveaux de prix pour savoir à quel niveau il n'y a plus d'excès de demande est une manière de trouver des points importants de la fonction de demande. Dans le cas d'une discrimination liée à un signal exogène,

l'identification des caractéristiques communes aux utilisateurs qui présentent une demande individuelle similaire peut aussi se faire par tâtonnement.

La méthode par tâtonnement peut donc servir de moyen de révélation de l'information. Cependant, cette solution implique une mise en place progressive de la tarification de pointe optimale.

Un autre moyen est de faire des estimations de la fonction de demande. Pour cela, il faut collecter de nombreuses données sur les compagnies. La constitution d'une base de données peut se faire en recensant différents éléments, tels que diverses données financières des compagnies ou certains de leurs choix stratégiques, observés lorsque plusieurs possibilités s'offrent à elles. Ce travail pourrait être complété par une enquête auprès des compagnies.

À partir de ces données, nous pourrions utiliser les outils économétriques afin de connaître les déterminants de la demande.

Sur un plan théorique, nous savons qu'un système d'enchères peut servir à faire révéler aux compagnies leurs disponibilités à payer. Le choix de règles adéquates de l'enchère, notamment la règle de paiement qui spécifie qui paye quoi, permet d'atteindre l'objectif recherché.

Cependant, la mise en place d'enchères, dans un contexte d'interdépendances, notamment des créneaux d'atterrissage et de décollage, et la satisfaction de certaines hypothèses indispensables à l'optimalité du résultat, reste complexe.

Conclusion

La présentation de quatre des redevances dues par les utilisateurs d'un aéroport pour l'atterrissage ou le décollage d'un avion nous a permis de voir quels éléments servaient à déterminer leur montant. Les redevances pour les aéroports de Paris, définies par ADP et l'autorité nationale, le sont dans un souci de financer les services, et non d'en allouer les ressources. Un choix a été fait de séparer la question de la tarification de celle de l'allocation. Jusqu'à présent le choix de la France, comme de la plupart des pays, pour les aéroports congestionnés a été celui d'un rationnement par les quantités, à travers un système d'allocation de créneaux aéroportuaires, ces « slots » étant gratuits.

Cependant, d'un point de vue économique cette séparation n'est pas optimale : la tarification et l'allocation doivent être traitées conjointement. Une manière de le faire est d'appliquer une tarification de pointe pour les diverses redevances aéroportuaires. Des prix de pointe viendraient remplacer la gratuité des slots et assureraient un rationnement efficace de la demande.

Le caractère aléatoire de la demande d'infrastructure et de prestations aéroportuaires ne fait quasiment pas de doute. C'est donc au moins une tarification de pointe reposant sur les disponibilités à payer qui doit s'appliquer. L'écart entre les coûts de l'opérateur en période de pointe et en période creuse peut aussi s'ajouter à l'écart entre les disponibilités à payer : cette partie des coûts s'expliquant alors, non par la congestion, mais par la volonté de faire en sorte que les coûts fixes d'installation d'une infrastructure liée à une plus forte demande à

certains moments soient reflétés dans les prix.

Face au problème de congestion, ce sont donc des redevances fondées sur les coûts de cette congestion imposés à autrui ou sur les propres disponibilités à payer des agents pour être servis à temps qui sont à envisager. Dans les deux cas, un approfondissement des connaissances de la demande est nécessaire. Les prochaines études devraient donc principalement s'orienter vers la constitution d'une base de données sur les compagnies et vers des estimations de la fonction de demande.

Résumé

L'objet de notre étude est la congestion aéroportuaire : nous cherchons à résorber ce problème, afin que l'offre aéroportuaire soit de meilleure qualité et ne pénalise pas la demande. Notre intérêt se porte donc sur les infrastructures aéroportuaires et les prestations qui y sont afférentes. Ces infrastructures sont essentiellement celles des installations au sol : pistes, voies de circulation, aires de stationnement, aérogares. Mais elles concernent aussi les infrastructures de navigation aérienne et les prestations de contrôle du trafic aérien. Tous les services apportés aux compagnies aériennes et aux passagers dans les aéroports ou aux abords donnent lieu à la perception de redevances.

Cette étude s'inscrit dans le cadre de travaux de recherche de la Direction de la Navigation Aérienne, de la Direction du Transport Aérien et de l'École Nationale de l'Aviation Civile. Face à une situation de congestion, elle vise à apporter à la DGAC des éléments d'appréciation sur la pertinence et la portée d'une tarification d'usage des infrastructures aéroportuaires à la pointe. En effet, La congestion aéroportuaire étant temporelle, du fait des fluctuations de la demande dans le temps, et l'accès aux infrastructures ne présentant pas la caractéristique de pouvoir être stocké, une tarification de pointe pourrait convenir.

La tarification de pointe est une solution, par exemple lorsqu'il existe une demande trop forte par rapport à la capacité d'une infrastructure. Mais, elle peut aussi être utile quand il n'existe pas de contrainte sur la capacité : étant donné que les coûts fixes d'installation d'une infrastructure nécessitent une élévation du prix au-dessus du coût marginal, la tarification de pointe peut remplir efficacement ce rôle. Cependant, il est vrai que si la capacité est contrainte, le problème est plus urgent et paraît par conséquent plus évident.

La première approche de ce travail est théorique ; elle est suivie d'une approche pratique. Dans un premier temps, nous nous appuyons sur les modèles théoriques et sur les résultats auxquels ils aboutissent, puis dans un deuxième temps, nous étudions les redevances aéroportuaires de pointe déjà appliquées et nous faisons une analyse de la situation française actuelle.

Dans le premier chapitre, nous présentons les déterminants fondamentaux des prix et leurs interactions. La maîtrise des coûts et le niveau de la demande jouent un rôle crucial dans la détermination des prix d'accès à une infrastructure.

Les interactions entre les coûts et les prix peuvent être de deux ordres. Dans le premier cas, la régulation « *cost-of-service* » repose sur la couverture totale des coûts. Des dépenses liées à l'infrastructure sont d'abord engagées et les prix sont ensuite déterminés de manière à ce que les recettes compensent les coûts. Par conséquent, c'est à partir des coûts que les

prix sont établis. Dans le second cas, la régulation « *price-cap* » repose sur un plafond mis à l'évolution des prix. Les prix sont d'abord calculés en fonction du taux de progression autorisé et l'opérateur doit ensuite engager des dépenses qui ne dépassent pas les recettes anticipées. Par conséquent, c'est à partir des prix que le niveau des coûts est décidé.

Les interactions entre la demande et les prix peuvent aussi être de deux ordres. Dans un premier cas, la discrimination tarifaire au second degré laisse les utilisateurs choisir une combinaison associant au prix un niveau d'utilisation de l'infrastructure ou un niveau de qualité du service. À travers les différents prix proposés, l'objectif de l'opérateur est d'influencer la demande, par exemple de manière à l'étaler dans le temps lorsque la capacité est atteinte. Ainsi, en proposant différents niveaux de prix variant avec la quantité ou la qualité, l'opérateur s'attend à une modification de la demande des utilisateurs. Dans un second cas, la discrimination tarifaire au troisième degré repose sur une différenciation des prix entre les utilisateurs de l'infrastructure sur la base de caractéristiques exogènes différentes entre ces agents. Les prix discriminants sont établis dans un souci d'accroître le bien-être, en rendant l'infrastructure accessible à un plus grand nombre d'utilisateurs, ou de maximiser le profit. C'est donc à partir de l'état de la demande, facilement segmentable et faible de la part de certaines catégories d'utilisateurs, que les prix sont fixés.

L'intérêt de discuter de la discrimination tarifaire est de lever certaines ambiguïtés fortes. En effet, il existe des *a priori* négatifs à l'encontre de la tarification de pointe, car elle est souvent associée à la discrimination tarifaire, qui elle-même est considérée comme portant préjudice aux agents économiques. Or, s'il est vrai que la tarification de pointe peut être discriminante, cela ne signifie pas qu'elle est injuste. La discrimination tarifaire peut, sous certaines conditions, être bonne sur le plan social.

Dans un deuxième chapitre, nous nous intéressons au dilemme auquel un gestionnaire d'une infrastructure, confronté à une demande fluctuant dans le temps, fait face lorsqu'il doit déterminer le niveau de sa capacité. Soit cette capacité est plus faible que le niveau maximal de la demande : l'opérateur ne sera alors pas en mesure de satisfaire à tout moment la demande. Soit cette capacité est suffisante pour satisfaire la demande de pointe : il aura alors d'importantes capacités excédentaires pendant les périodes creuses. Une tarification de pointe peut contribuer à atténuer ce problème.

Lorsque la demande est déterministe, il est possible de toujours la satisfaire, même si elle est très élevée. Les prix peuvent alors reposer sur les coûts de l'opérateur. L'intérêt d'une tarification de pointe, dans ce cas, est de faire payer un prix plus élevé aux utilisateurs de l'infrastructure pour lesquels le sur-dimensionnement a été mis en place. La justification de ce système de prix reposerait sur la base du principe du « coût du service » : l'utilisateur devant payer les coûts qu'il engendre. Mais, même si les coûts et les prix sont liés, de la discrimination tarifaire peut se justifier, avec des demandes de pointe et de hors pointe très proches. La demande des périodes creuses doit alors contribuer à financer la capacité des périodes de pointe.

Dans l'autre situation, celle d'une demande aléatoire, toute la demande ne peut être satisfaite, elle doit être rationnée. Mis à part un rationnement aléatoire, le rationnement passe par une tarification de pointe reposant sur la base du principe de la « valeur du service ».

L'intérêt, ici, est de faire prendre conscience aux utilisateurs du sur-coût qu'ils font supporter aux autres, en faisant en sorte que les externalités soient intégrées dans les calculs individuels. Un autre intérêt est de faire payer aux utilisateurs la valeur qu'ils accordent à l'infrastructure. Seul le choix entre un rationnement selon l'ordre croissant des disponibilités à payer et selon leur ordre décroissant est discrétionnaire, lorsque la demande est aléatoire.

Ainsi, une tarification de pointe n'aboutit quasiment jamais à des prix où chacun paie une contribution en proportion de la capacité, des coûts physiques, qu'il impose. Soit même les utilisateurs de la période creuse payent une partie des coûts, soit les coûts de congestion sont inclus dans les coûts, soit il faut discriminer entre les agents pour les inciter à déplacer leur demande.

Dans un troisième chapitre, nous étudions les modulations tarifaires horaires pratiquées par de nombreux aéroports et dans d'autres secteurs.

L'industrie de réseau dont l'organisation se rapproche le plus de celle de l'aérien est le système ferroviaire. En France, en 1997, la SNCF a été divisée en deux entreprises. RFF est devenu un « fournisseur » de la SNCF et perçoit à ce titre des redevances pour l'utilisation de l'infrastructure qu'il met à la disposition de la SNCF. RFF a adopté des redevances pour l'utilisation du réseau modulables selon les heures.

Concernant les aéroports, les principales tranches horaires sur-tarifées sont la nuit et les heures en dehors des périodes normales d'activité de l'aéroport. Ces modulations tarifaires n'ont donc rien de commun avec la tarification de pointe. Les modulations tarifaires liées aux pointes sont en revanche beaucoup moins pratiquées. En matière de redevances aéroportuaires, comme dans beaucoup d'autres domaines, les anglo-saxons ont été les précurseurs d'une modulation horaire fonction de la pointe. Les aéroports new-yorkais en 1968, puis les aéroports londoniens en 1972 ont décidé de faire dépendre leurs redevances de l'heure de l'atterrissage ou du décollage, en raison de problèmes de congestion. À partir des expériences étrangères en matière de redevances aéroportuaires de pointe, on voit quelles adaptations des résultats théoriques peuvent être faites.

Les redevances londoniennes sont croissantes avec le degré de la pointe. La progression de ces redevances permet de se rapprocher d'un péage optimal, qui définit un prix pour chaque instant. Cet étalement des redevances permet aussi d'éviter des variations trop brutales de la période creuse à la période de pointe. Avec un palier intermédiaire pour les pointes modérées, les effets de seuil sont limités.

Les compagnies sont faiblement sensibles aux variations des redevances. On observe que les modulations tarifaires entre les périodes doivent être importantes pour induire un changement de comportement des compagnies. On observe que de 2002 à 2003, les aéroports londoniens ont opéré un large réajustement à la hausse de leurs redevances afin d'accroître l'impact de la surcharge.

Certains aéroports appliquent une tarification de pointe qui touche indirectement l'aviation générale. D'autres ont choisi des modulations de tarif qui visent directement le transport de marchandises ou les opérations non commerciales. Les premières années de la mise en œuvre de la tarification de pointe pour les aéroports londoniens ont permis d'observer son impact sur le trafic. Ce sont surtout les opérations non-commerciales ou de transport

de marchandises qui furent affectées par la surcharge en raison de l'importance des coûts supplémentaires qu'elle a représentée. À Manchester, les mesures visant à réduire la congestion, en période de pointe, ont pour cible avant tout les avions cargo. Les aéroports new-yorkais confrontés à un sérieux problème de congestion souhaitaient réduire le trafic de l'aviation générale. Ils ont donc instauré une taxe pour les atterrissages et les décollages aux heures de pointe pour les avions de moins de 25 places. L'aéroport de Toronto au Canada vise directement l'aviation générale : il surtaxe les appareils légers aux périodes de pointe.

Par ailleurs, la pratique dans les trois principaux aéroports de la région de New-York, est proche des recommandations de BRUECKNER, qui consiste à corréler négativement les redevances d'une compagnie à son pouvoir de marché dans un aéroport. Les compagnies aériennes et les vols qui relient quotidiennement l'aéroport voient leurs charges réduites par rapport aux autres. Les opérateurs réguliers profitent donc de réductions.

Les effets d'une tarification de pointe sur le trafic restent difficiles à observer, étant donné que ces aéroports congestionnés attribuent des slots, situation qui empêche de connaître la véritable demande des compagnies.

Dans le dernier chapitre, la présentation de quatre des redevances dues par les utilisateurs d'un aéroport pour l'atterrissage ou le décollage d'un avion permet de voir quels éléments servent à déterminer leur montant. Les redevances pour les aéroports de Paris, définies par ADP et l'autorité nationale, le sont dans un souci de financer les services, et non d'en allouer les ressources. Jusqu'à présent le choix de la France, comme de la plupart des pays, pour les aéroports congestionnés a été celui d'un rationnement par les quantités, à travers un système d'allocation de créneaux aéroportuaires, ces « slots » étant gratuits.

Le caractère aléatoire de la demande d'infrastructure et de prestations aéroportuaires ne fait quasiment pas de doute. Face au problème de congestion, ce sont donc des redevances fondées sur les coûts de cette congestion imposés à autrui ou sur les propres disponibilités à payer des agents pour être servis à temps qui sont à envisager. Dans les deux cas, un approfondissement des connaissances de la demande est nécessaire. Les prochaines études devraient donc principalement s'orienter vers la constitution d'une base de données sur les compagnies et vers des estimations de la fonction de demande.

Le problème qui demeure est celui de l'information. À notre époque, la détention d'information est source de richesse. À l'inverse, l'absence d'information est difficilement surmontable, car l'acquisition est par conséquent coûteuse. Des moyens doivent être mis en place pour obtenir une information pertinente et à moindre coûts.

Cette étude nous enseigne que tant que la capacité est suffisante, la tarification doit reposer sur une approche par les coûts, situation qui n'exclue pas une tarification de pointe. En revanche, dès que la demande dépasse l'offre, notamment parce qu'elle fluctue de façon aléatoire, l'approche pour la tarification doit alors être celle de la demande. Dans ce cas, il est indispensable de lier la question de la tarification de pointe à celle de l'allocation des créneaux. Tant que le principe qui prévaudra pour l'allocation des créneaux sera celui des droits du grand-père, l'efficacité de la tarification de pointe sera limitée.

Bibliographie

[1] R. ARNOTT, A. DE PALMA et R. LINDSEY. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review*, 83(1):161–179, mars 1993.

[2] M. BOITEUX. Sur la gestion des monopoles astreints à l'équilibre budgétaire. *Econometrica*, 24:22–40, 1956.

[3] J.K. BRUECKNER. Internalization of airport congestion. *Air Tranport Management*, 8:141–147, 2002.

[4] G. BRUNEKREFFT. Price capping and peak-load pricing in network industies. *Working Paper*, decembre 2000.

[5] D. BÖS. *Hanbook of public economics*, chapitre *Public sector pricing*, volume 1, pages 129–211. Elsevier Science Publishers B.V., A.J. AUERBACH et m. FELDSTEIN edition, 1985.

[6] J. CALZADA. Capacity charge and peak-load pricing. *Working Paper*, janvier 2003.

[7] M.A. CREW, C.S. FERNANDO et P.R. KLEINDORFER. The theory of peak-load pricing : A survey. *Journal of Regulatory Economics*, 8:215–248, 1995.

[8] N. CURIEN. *Économie des réseaux*. La Découverte, coll. « Repères », 2000.

[9] M. KATZ. Non uniform pricing, output and welfare under monopoly. *Review of Economic Studies*, 50(1):37–56, 1983.

[10] J-J. LAFFONT et J. TIROLE. *A Theory of Incentives in Procurement and Regulation*. MIT Press, 1993.

[11] M.G. MARCHAND. Priority pricing. *Management Science*, 20(7):1131–1140, mars 1974.

[12] J.M. MIRRLEES. An exploration in the theory of optimum income taxation. *Review of Economic Studies*, 38:175–208, 1971.

[13] S.A. MORRISON. The equity and efficiency of runway pricing. *Journal of Public Economics*, 34:45–60, 1987.

[14] L. PHLIPS. *The economics of price discrimination*. Cambridge University Press, 1983.

[15] A.C. PIGOU. The economics of welfare. *London: Mc Millan*, 1920.

[16] M. RAFFARIN. *Le contrôle aérien en France : congestion et mécanismes de prix*. Thèse de Doctorat, Université de Paris-I – Panthéon-Sorbonne, 2002.

[17] J. ROBINSON. *The Economics of Imperfect Competition*. London: Mc Millan, 1933.

[18] S. SCHMALENSEE. Output and welfare implications of monopolistic third-degree price discrimination. *American Economic Review*, 71:242–247, 1981.

[19] P.O. STEINER. Peak loads and efficient pricing. *Quarterly Journal of Economics*, 71:585–610, novembre 1957.

[20] J. TIROLE. *The Theory of Industrial Organization*. MIT Press, 1989.

[21] W. VICKREY. Congestion theory and transport investment. *American Economic review*, 59:251–261, mai 1969.

Table des matières