INRETS INSTITUT NATIONAL DE RECHERCHE SUR LES TRANSPORTS ET LEUR SÉCURITÉ

Véronique SAUVADET

« Prévision d'ensemble »
Usage d'un concept employé en météorologie dans le domaine du trafic routier

Rapport de stage

Octobre 2001

INSTITUT NATIONAL DE RECHERCHE
SUR LES TRANSPORTS ET LEUR SÉCURITÉ

Véronique SAUVADET
Institut de Statistique de l'Université Pierre et Marie Curie (ISUP)
Filière Industrie et Services

« Prévision d'ensemble »
Usage d'un concept employé en météorologie dans le domaine du trafic routier

Rapport de stage

Octobre 2001

Maître de stage
Mehdi DANECH-PAJOUH
INRETS-GRETIA

Remerciements

Mes remerciements vont en premier lieu à Monsieur BOSQ et Monsieur DELECROIX pour avoir accepté le sujet proposé et ainsi permis la réalisation de ce stage. Leurs conseils m'ont été une aide précieuse.

Je tiens également à remercier tout particulièrement Monsieur DANECH-PAJOUH, Chargé de Recherche à l'Institut National de Recherche pour les Transports et leur Sécurité. Son accueil, son soutien, sa confiance, sa disponibilité sans faille mais surtout sa connaissance statistique ainsi que son expérience dans le domaine du trafic routier m'ont permis de progresser tant d'un point de vue personnel que professionnel.

Mes remerciements s'adressent aussi à Monsieur ROUX météorologue, Directeur de Recherche au CNRS, qui a accepté de lire la partie de ce rapport concernant les méthodes de prévision en météorologie. Ses conseils, toujours judicieux, m'ont permis de mieux appréhender et de comprendre les mystères du temps.

Enfin, ma reconnaissance va à toute l'équipe du département GRETIA de l'INRETS pour sa bonne humeur et sa coopération. Son accueil chaleureux a été un soutien essentiel pour l'accomplissement de ce travail.

A toutes et à tous, un grand merci …….

Sommaire

Introduction générale

Dans le cadre de sa collaboration avec l'ISUP, l'INRETS (Institut de Recherche sur les transports et leur sécurité) a proposé un sujet ayant comme objet : l'ensemble des techniques d'évaluation à priori. Ce travail a été réalisé au sein du laboratoire GRETIA (Génie des Réseaux de Transports et Informatique Avancée).

L'objectif de cette étude est de vérifier si les techniques d'évaluation à priori des prévisions utilisées en météorologie peuvent être applicables dans le domaine des transports.

Les prévisions météorologiques ne sont pas parfaitement exactes et ne le seront jamais. Du caractère chaotique de l'écoulement atmosphérique et d'une petite incertitude sur l'état actuel de l'atmosphère résulte rapidement une grande incertitude sur son état futur. Il apparaît donc souhaitable, pour beaucoup d'applications, d'accompagner la prévision d'une estimation à priori de l'incertitude correspondante. Pour ce faire, il est nécessaire de développer un système de prévision de l'incertitude qui prenne en compte les caractères spécifiques de l'état actuel de l'atmosphère (distribution spatiale des observations disponibles, état d'instabilité de l'écoulement…). Sous sa forme la plus achevée, ce système produira des probabilités quantitatives pour l'occurrence d'événements spécifiques.

L'application de manière exacte des concepts météorologiques au trafic semble difficile, celui-ci étant nettement moins sensible à l'état initial. L'état du trafic sur une route dépend de variables telles que la route, l'environnement, les véhicules et les conducteurs. Les prévisions de l'état du trafic (à l'horizon variant de quelques jours à un an) peuvent se faire selon différentes conditions de l'offre routière (conditions météorologiques, état de la route, disponibilité des autres modes de transport …) et de la demande de déplacements (événements prévisibles et non prévisibles).

La première partie est la synthèse d'une recherche bibliographique donnant un aperçu des méthodes de prévisions météorologiques. L'objectif de cette recherche n'était pas d'analyser tous les concepts météorologiques très complexes mais de comprendre la philosophie des méthodologies utilisées dans le cas d'une prévision du temps. Ainsi, sont présentées les observations disponibles en météorologie ainsi que les modèles numériques utilisés. Ces deux notions permettent la compréhension de l'origine de l'incertitude et amènent à considérer la prévision d'ensemble qui permet l'obtention, pour une échéance donnée, non pas d'une seule valeur mais plusieurs. On aboutit alors à la notion de prévision probabiliste.

La deuxième partie quant à elle porte sur l'utilisation des concepts météorologiques dans le cas du trafic routier. Cette application s'est faite sur un modèle de prévisions du dispositif Bison Futé. Ce dispositif, bien connu de tous, propose des prévisions du trafic journalier (ou horaire) un an à l'avance. La nature des données à prédire et le peu d'informations disponibles à aussi long terme donnent au modèle une complexité non négligeable. Dans cette partie nous présentons d'abord le modèle utilisé par le logiciel Bison Futé, produit d'une étude réalisée par le SETRA, le CETE de la région Nord-Picardie et l'INRETS. Ce modèle a la structure classique d'un modèle d'analyse de variance puisque les paramètres explicatifs qui sont utilisés pour prévoir le trafic sont des variables qualitatives caractérisant calendairement les jours considérés. Après avoir réalisé une rapide analyse à posteriori des prévisions, nous avons tenté d'utiliser le concept de système du pauvre pour permettre l'obtention d'un ensemble de prévisions pour chaque jour. L'analyse de cet

ensemble peut conduire à la création d'un indice de confiance et à une possible formulation probabiliste.

La troisième partie propose l'application du concept d'évaluation a priori des résultats provenant des outils de simulation du trafic.

Les modèles permettent de donner une représentation simplifiée de la réalité sous forme de lois (c'est-à-dire de variables et de relations entre ces variables) et sont destinés aussi bien à améliorer la connaissance de cette réalité qu'à être partie intégrante d'un processus de contrôle. L'objectif des modèles est alors de tester des hypothèses d'évolution, d'évaluer des stratégies de commande ou l'influence d'un paramètre sur le comportement d'ensemble. La simulation est un processus de résolution du modèle, c'est-à-dire le calcul des états successifs. Un même modèle peut faire l'objet de divers modes de résolution, et un outil de simulation intègre éventuellement plusieurs modèles.

A partir d'un ensemble de conditions initiales (état du réseau au début de l'étude) et de conditions aux limites (demande en entrée du réseau, contraintes en sortie du réseau, incidents…) le modèle doit permettre de déterminer l'évolution des variables. Il existe essentiellement trois grandes catégories de modèles en trafic :

- Modèles microscopiques
- Modèles macroscopiques
- Modèles mésoscopiques

Chaque modèle utilise des lois, qui sont issues, pour la plus grande partie, de la mécanique des fluides.

La simulation est fondée sur la discrétisation des variables de temps et d'espace. On calcule donc l'état du système à un instant t à partir de son instant à $t-\Delta t$, c'est la simulation dite « pas à pas ». Lorsque l'espace est discrétisé, certaines variables sont moyennées sur les zones comprises entre x et $x+\Delta x$ (par exemple, dans les modèles macroscopiques, la densité est supposée constante sur chaque section).

Ces méthodes de résolutions ainsi que les équations utilisées forment la première analogie avec la météorologie. De plus, même si le trafic est moins sensible, il existe une incertitude sur les données initiales qui peut avoir une répercussion sur l'évolution future du trafic. La méconnaissance des données initiales influe donc sur les résultats de la simulation. Il importe alors d'avoir une analyse critique des résultats par rapport aux données d'entrée. Ceci peut permettre de juger, qualitativement ou quantitativement, de la validité et de la portée des résultats. Nous verrons donc comment le concept de prévision d'ensemble peut s'appliquer et quelle analyse peut aboutir à une évaluation a priori du résultat. Il convient de noter que cette méthode d'évaluation, telle qu'elle sera présentée, ne concerne que les résultats issus de la simulation. Elle est, en effet, totalement indépendante des phases de calibrage et de validation du modèle.

Première partie :
Un aperçu des méthodes
de prévisions météorologiques

L'objectif principal de la météorologie est la construction des modèles permettant la prévision du temps à différentes échéances (immédiate inférieure à 6 heures, courte inférieure à 3 jours, moyenne entre 3 et 7 jours). L'objectif de Météo France est de faire des prévisions, celui des chercheurs est la compréhension des phénomènes observés. Le modèle est alors un outil comme le sont les observations. Les lois physiques qui régissent l'écoulement de l'atmosphère sont bien connues (il reste encore des incertitudes, par ex : turbulence, nuages, rayonnement….), néanmoins la prévision météorologique reste toujours très difficile. En effet, aussi précise que soit la connaissance de l'état présent de l'atmosphère, il reste malgré tout des incertitudes inévitables. Et celles-ci se transforment vite en une grande incertitude sur l'état futur.

L'évolution constante dans la recherche d'une prévision toujours plus précise est d'abord passée par l'augmentation de capacité et de rapidité des calculateurs puis aujourd'hui par l'amélioration de la connaissance de l'état initial. Une prévision plus précise nécessite en effet à la fois une augmentation de la puissance de calculs (on est encore loin d'avoir atteint les limites) et une amélioration des observations. Malgré l'importance croissante de l'assimilation des données météorologiques pour la détermination de l'état initial, les approches aujourd'hui consistent à ne plus s'en tenir à une unique prévision déterministe. **Au-delà de la prévision, on cherche aussi à prévoir l'incertitude de l'état futur de l'atmosphère.**

L'élaboration d'une prévision du temps est un processus complexe, que l'on peut cependant décomposer en trois étapes :

- une analyse de la situation atmosphérique présente (repérer des phénomènes météorologiques)
- une prévision de l'évolution de ces phénomènes
- la prévision du temps sensible

Après une présentation succincte des observations disponibles en météorologie qui met en évidence l'existence de l'incertitude sur l'état initial, est expliquée la modélisation. Le principe de construction du modèle de prévisions numériques permet de comprendre comment se déduit la prévision du temps sensible, c'est-à-dire celle d'un phénomène météorologique.

La notion d'incertitude de l'état futur nous amène à parler de la prévision d'ensemble pour en venir à la prévision probabiliste. On ne prévoit plus une valeur d'un paramètre mais une loi de probabilité suivie par ce paramètre.

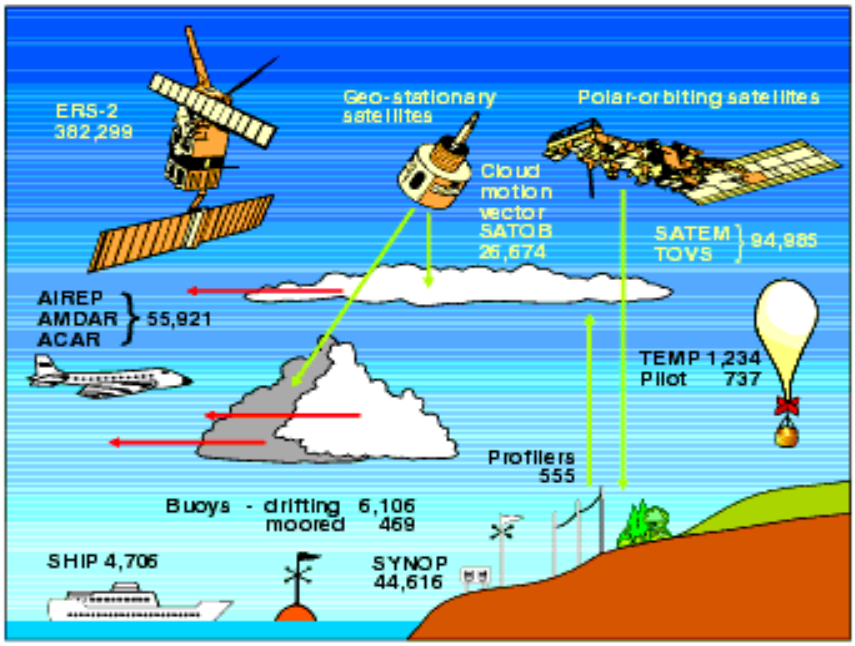
Enfin, la méconnaissance des observations conduit à utiliser le principe de l'assimilation de données. Ce concept permet de combiner l'état observé et l'état prévu pour en déduire l'état le plus probable de l'atmosphère.

I. Les observations en météorologie

1) Méthodes de recueil des données :

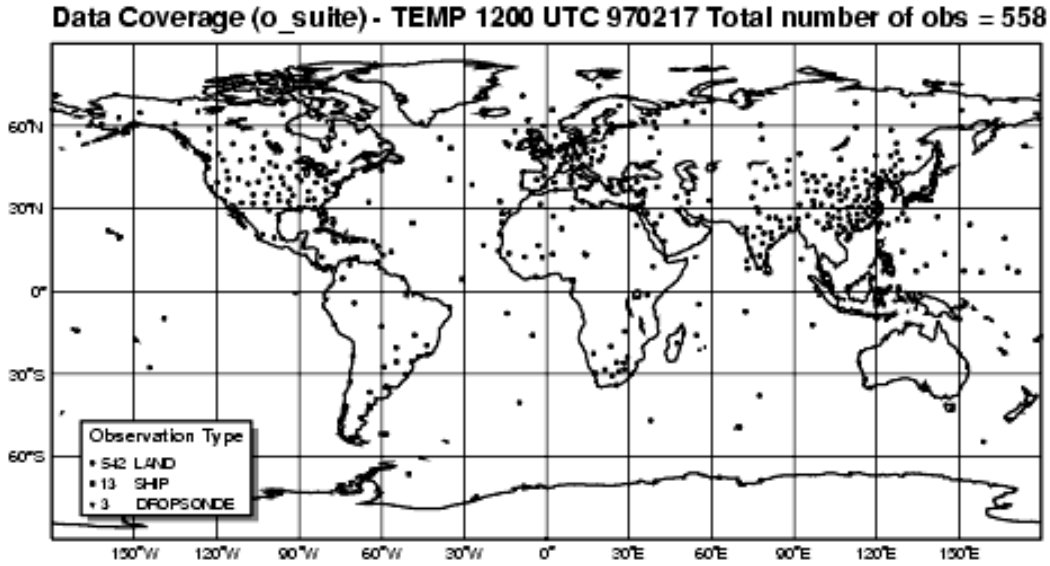
Les modèles de prévision ne peuvent remplir efficacement leur mission qu'à la condition d'être en permanence nourris par un flux d'informations sur l'état de l'atmosphère. Pour cela, l'Organisation Météorologique Mondiale gère un réseau international de télécommunications qui rend disponible l'ensemble des observations effectuées **dans le monde entier.**

Pour réaliser l'observation de l'état de l'atmosphère, les services météorologiques du monde entier effectuent des mesures à des heures fixes. Deux fois par jour, à 0 h et 12h UTC, environ sept cents **ballons** sont lancés simultanément dans le monde. Equipés d'instruments ils mesurent en continu la température, l'humidité et la pression et les transmettent par radio au sol. Du déplacement des ballons, on peut déduire le vent, en les suivant avec des radars ou en utilisant des radionavigations pour connaître leur position. Ces ballons s'élèvent à des altitudes variables, entre 20 et 30 km, avant d'éclater et de retomber sur terre ou en mer. Des mesures à bord d'**avions de ligne** permettent également d'obtenir ces mêmes informations en altitude. D'autres mesures (plus nombreuses) sont effectuées au sol **dans les stations météorologiques, sur des navires, à partir de bouées**. Elles sont également complétées par des mesures effectuées à partir de **satellites météorologiques**. Ceux-ci en orbite basse (900km) effectuent des mesures qui permettent d'estimer la répartition verticale de température et d'humidité dans les régions où les radiosondages sont rares. Ceux placés sur orbite géostationnaire à 36 000 km ont un champ de vue plus large. Les images qu'ils enregistrent et transmettent au sol fournissent des indications sur le développement des nuages et sur les vents.



2) Problèmes liés aux observations :

De part la description des méthodes de recueil des données météorologiques, on imagine assez bien les difficultés que présentent les observations.
Dans la situation physique idéale, on s'attend à ce que les observations précisent sans aucune ambiguïté la valeur des paramètres d'état au temps initial t=0. Ce cas idéal n'existe pas en météorologie pour diverses raisons. En effet, **connaître parfaitement l'état de l'atmosphère à un instant donné signifie une connaissance en tout point avec une précision infinie**.



Tout d'abord comme le montre la carte, **les observations sont très mal réparties autour de la planète** (mesures inexistantes ou presque sur les océans et hémisphère sud). La description précédente des données météorologiques montre aussi que le nombre de mesures dont on dispose surtout en altitude, est extrêmement faible alors que les modèles utilisés font des prévisions sur des milliers de points.

Les mesures par radiosondage (observations à l'aide d'un ensemble d'instruments transportés par ballon) sont simultanées dans toute la planète. Mais la plupart des autres **mesures ne sont pas simultanées ni dans l'espace, ni dans le temps. Elles ne définissent pas l'état de l'atmosphère à un instant précis**. (Même en supposant le modèle discrétisé on ne dispose d'aucun moyen pour mesurer l'état de l'atmosphère à un instant donné en tous les points de discrétisation).

Il existe également des **erreurs liées à la mesure** proprement dite (erreur matérielle instrumentale). Enfin toutes les mesures sont mises en forme et codées pour pouvoir être transmises sur les lignes du système mondial de télécommunications météorologiques. Il existe **des procédés sophistiqués pour décoder les messages et ainsi restituer les paramètres météorologiques**. Ils ont des déficiences et **causent des erreurs importantes** qu'il est nécessaire de prendre en compte, erreurs qui restent néanmoins négligeables par rapport aux autres sources d'erreurs.

On peut en conclure que les séries de données disponibles pour représenter éventuellement le point de départ d'une prévision sont incomplètes et inexactes et ne suffisent pas à fournir, par elles-mêmes, une description adéquate de l'atmosphère.

II. Modèle de prévisions numériques

Le moyen de prévision le plus sûr et le plus efficace est aujourd'hui **la modélisation numérique c'est-à-dire le calcul de proche en proche**, à partir de l'état observé de l'atmosphère, de l'évolution future de l'écoulement. La réalisation d'une prévision sera décomposée en deux étapes : l'observation de l'état de l'atmosphère à un instant donné et le calcul de ses états successifs.

1) Principe :

La prévision numérique consiste en l'intégration temporelle d'un système d'équations aux dérivées partielles décrivant les grandes lois physiques de l'atmosphère. Ces équations, impossibles à résoudre analytiquement, sont intégrées par des méthodes numériques approchées, à partir d'un état initial reflétant le mieux possible la description présente de l'atmosphère. L'objectif est donc de déterminer les valeurs futures des grandeurs caractéristiques de l'atmosphère en partant de valeurs initiales connues grâce aux observations météorologiques.

La construction d'un modèle numérique comprend deux étapes :

- la première consiste à établir un système d'équations
- la seconde, dite de « numérisation », consiste à remplacer les équations portant sur des variables continues par des équations portant sur des variables discrètes et dont les solutions sont obtenues au moyen d'un algorithme approprié. La numérisation des équations peut être effectuée en utilisant un modèle en points de grille. Le modèle en points de grille correspond à la division de l'espace au moyen de boîtes définies par une grille horizontale et par un nombre de niveaux verticaux. Dans chaque boîte l'atmosphère est supposée homogène. Plus la taille des boîtes est faible, meilleure est la description des paramètres du modèle représentant l'état de l'atmosphère.

  Par exemple, en 2000, le modèle utilisé par le CEPMMT (Centre Européen pour les Prévisions Météorologiques à Moyen Terme) est construit sur un maillage dont la résolution est d'environ 40 km dans la direction horizontale, avec 60 niveaux verticaux inégalement répartis. Il faut, dans ces conditions, 25 millions de paramètres (température, composantes du vent…) pour décrire l'état de l'écoulement à un instant donné à tous les points du maillage sur la totalité de l'atmosphère.

2) Equations primitives :

Pour l'étude de la dynamique de l'atmosphère on utilise les équations primitives (équations d'évolution d'un fluide en équilibre hydrostatique, lois de la thermodynamique et de la dynamique des fluides). Néanmoins il apparaît obligatoire d'apporter des hypothèses simplificatrices telle que l'approximation hydrostatique qui suppose que l'accélération verticale de l'air est négligeable. Dans ces équations, les variables indépendantes sont les coordonnées horizontales, x et y et la pression p que l'approximation hydrostatique permet d'utiliser comme coordonnée verticale.

$$(1)\quad \frac{\partial \vec{U}}{\partial t} = -(\vec{U}.\vec{\nabla})\vec{U} - \omega\frac{\partial \vec{U}}{\partial p} - f\,\vec{k}\wedge\vec{U} - g\nabla z + \vec{F}$$

$$(2)\quad \frac{\partial T}{\partial t} = -\vec{U}.\vec{\nabla}T - \omega\frac{\partial T}{\partial p} + \frac{RT\omega}{C_p p} + \frac{Q}{C_p}$$

$$(3)\quad \frac{\partial q}{\partial t} = -\vec{U}.\vec{\nabla}q - \omega\frac{\partial q}{\partial p} + g\frac{\partial E}{\partial q} \quad \text{si } q < q_s \text{ ou } \frac{dq_s}{dt} > 0$$

$$(3\text{bis})\quad \frac{\partial q}{\partial t} = -\vec{U}.\vec{\nabla}q - \omega\frac{\partial q}{\partial p} + \frac{dq_s}{dt} \quad \text{si } q = q_s \text{ ou } \frac{dq_s}{dt} < 0$$

Ces trois équations définissent en tout point, à un instant donné, les dérivées par rapport au temps des champs de vitesse horizontale $\vec{U}$, de température T et d'humidité q en fonction des valeurs de ces champs à l'instant considéré.

(4) $\frac{\partial \omega}{\partial t} + \vec{\nabla} . \vec{U} = 0$

(5) $\frac{\partial z}{\partial p} + \frac{RT}{gp} = 0$

Ces deux équations définissent en fonction des variables présentes la « vitesse verticale » $\omega = \frac{dp}{dt}$ et l'altitude z.

$\vec{U}$: vecteur vitesse horizontale
T : température
q : humidité spécifique
p : pression
« vitesse verticale » $\omega = \frac{dp}{dt}$
altitude z
$\vec{\nabla}$: opérateur gradient horizontal

le symbole d représente une dérivée par rapport au temps prise en suivant une particule en mouvement. Cette dérivation (dérivation temporelle lagrangienne) s'oppose à la dérivée temporelle eurélienne notée ∂ et prise en un point fixe au cours du temps.
On obtient donc les deux expressions suivantes, pour une variable X(x, t) :

$$\frac{\partial X}{\partial t} = \frac{X(x_0, t+\Delta t) - X(x_0, t)}{\Delta t}$$

$$\frac{dX}{dt} = \frac{X(x_0, t+\Delta t) - X(x_i, t)}{\Delta t}$$ où x_i est la position à t de X arrivant à x_0 à $t+\Delta t$

On a la relation suivante entre les deux dérivées :

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \vec{U} . \vec{\nabla} + \frac{\partial}{\partial p}$$

La dérivée lagrangienne est donc la somme de la dérivée eurélienne avec des termes complémentaires appelés termes d'advection que l'on retrouve dans les équations précédentes.

- Dans la première équation (1) :

le terme $f\ \vec{k} \wedge \vec{U}$ représente la force de Coriolis ($f=2\Omega\sin\varphi$ où Ω est la vitesse de rotation de la Terre et φ la latitude ; $\vec{k}$ est le vecteur unitaire dirigé vers le haut).

$g\vec{\nabla}z$ représente la force horizontale de pression (g est l'accélération de la gravité).
$\vec{F}$ est la force de frottement horizontale exercée sur une unité de masse du fluide.

- Dans la deuxième équation (2) :

$\frac{R}{C_p}\frac{T\omega}{p}$ représente la variation de température due à une compression (ou détente) adiabatique lors de mouvements verticaux (R constante des gaz parfaits, C_p chaleur spécifique de l'air à pression constante). $\frac{Q}{C_p}$ représente la variation de température due à l'apport d'énergie extérieure (Q quantité de chaleur reçue par unité de temps par unité de masse du fluide).

- Dans la troisième équation (constituée des deux équations (3) et (3bis)) :

q_s représente l'humidité spécifique saturante.
E est le transport vertical turbulent de vapeur d'eau par unité de surface horizontale et par unité de temps.
Si l'air n'est pas saturé la variation de q est déterminée par l'advection et le flux turbulent de vapeur d'eau. Si l'air est et reste saturé, la variation d'humidité est égale à la variation d'humidité saturante.

- La quatrième équation (4) (équation de continuité) exprime la conservation de la masse.

Elle permet de calculer ω en tout point de l'atmosphère. (On prend en général ω=0 au niveau p=0, bien qu'il soit aberrant physiquement de supposer l'existence d'un niveau où la pression serait nulle). On prend ω=0 à la surface.

- La cinquième équation (5) exprime l'approximation hydrostatique.

En écrivant qu'au niveau du sol la composante normale de la vitesse du vent est nulle, on obtient la pression au sol, ce qui revient à connaître l'altitude sur un niveau de pression de référence.

Remarque : certains termes nécessaires à l'intégration des équations ne sont pas définis explicitement ($\vec{F}$, Q, E, q_s). Il est donc nécessaire d'en connaître une représentation paramétrique en fonction des différentes variables.

Ces équations primitives définissent un problème aux conditions initiales : à partir de la connaissance des valeurs des différents champs (vitesse horizontale, température, humidité) à un instant donné, les équations primitives permettent de calculer, au moins théoriquement, leur évolution future.
Les valeurs initiales des paramètres sont donc l'entrée du modèle (valeurs obtenues à partir des observations).
Les équations dont l'évolution temporelle est connue sont intégrées à chaque pas de temps de façon à prévoir sur la grille ces mêmes paramètres. **Ces champs obtenus à l'issue d'un certain nombre de pas de temps sont les sorties du modèle** pour une échéance de prévision donnée.

Les modèles numériques ne sont que des systèmes de prévision des grandeurs météorologiques et non véritablement de prévision du temps.

Pour une prévision locale, l'idée est d'établir un **lien statistique entre la prévision d'un modèle dynamique et le phénomène (ou paramètre) météorologique que l'on veut prévoir**. Le modèle dynamique calcule une prévision générale (sur une grille horizontale – tout ou une partie du globe- et verticale –à différentes altitudes-) des paramètres physiques pour une certaine échéance. **Les outils statistiques permettent de la transformer en une prévision locale** (un paramètre du temps sensible : température à 2m, vent à 10m, la durée d'insolation, l'occurrence de pluie, de neige…). Lors de telles études les prédicteurs sont constitués d'extraits de champs de paramètres prévus par le modèle dynamique considéré.

L'objectif des modèles dynamique est donc de déterminer les valeurs futures des grandeurs caractéristiques de l'atmosphère en partant de valeurs initiales connues grâce aux observations météorologiques.
Néanmoins il reste toujours une incertitude quant aux données initiales et donc une incertitude sur la prévision qui en découle.

III. La Prévision d'ensemble

En dépit de l'amélioration des systèmes de prévisions météorologiques réalisées à l'aide de modèles numériques, il s'avère impossible de fournir des prévisions précises au-delà d'une certaine limite. En effet comme l'atmosphère ne pourra jamais être complètement observée, le modèle dynamique de prévisions numériques du temps continuera toujours de calculer des prévisions à partir d'un état légèrement différent de l'état réel de l'atmosphère. Or, des prévisions à partir de différentes conditions initiales très similaires les unes des autres, peuvent aboutir à des résultats très différents dans le futur. De plus il perdure aussi des incertitudes sur la discrétisation spatio-temporelle et sur les paramétrisations.

Aussi, au lieu d'effectuer une prévision unique à partir d'une condition initiale, on effectue un ensemble de prévisions à partir d'un ensemble de conditions initiales légèrement perturbées. L'ensemble des conditions initiales doit représenter l'incertitude sur l'état présent de l'atmosphère, la dispersion des prévisions sera alors prise comme une mesure de l'incertitude sur l'état futur. Un tel système pourra définir des limites à l'intérieur desquelles on peut affirmer que se trouvent les valeurs des quantités météorologiques.
On peut donc appliquer le modèle de prévision numérique pour chaque état de l'atmosphère appartenant à l'ensemble des conditions initiales ce qui permet ainsi d'obtenir un ensemble de prévision. **A une condition initiale correspond une prévision numérique**.

1) Choix des conditions initiales :

Le problème fondamental de la prévision d'ensemble est d'effectuer un choix judicieux des situations initiales de façon à pouvoir obtenir un maximum de solutions éloignées les unes des autres avec un minimum d'états initiaux.

En effet, la taille de l'ensemble des conditions initiales est limitée par le nombre de calculs à effectuer pour obtenir la prévision d'ensemble. Les perturbations de départ qui servent à modifier légèrement l'état initial sont donc à choisir le plus judicieusement possible.
L'amplitude de ces perturbations doit toujours rester compatible avec les incertitudes portant sur la connaissance de l'état initial réel. **Chaque état initial utilisé comme point de départ**

d'une simulation représente un état possible de l'atmosphère, compatible avec les observations disponibles. On évitera de choisir un état initial complètement aberrant par rapport à la connaissance imparfaite que l'on a de l'atmosphère au travers des observations disponibles.

La prévision d'ensemble demande donc la construction d'un ensemble initial considéré comme un échantillon d'une loi de probabilité représentant l'incertitude. Or visualiser les distributions de probabilités de l'état initial est très difficile surtout lorsqu'elles nécessitent des lois de probabilité jointes se rapportant à un grand nombre de variables. On peut néanmoins utiliser la représentation de la trajectoire dans l'**espace des phases** (un espace des phases est une représentation géométrique d'un système dynamique, où chaque coordonnée des axes se rapporte à une des variables du système). Chaque point dans cet espace représente l'état du système à un instant donné. Les espaces des phases des modèles numériques de prévision du temps ont des millions pour dimension, chacune correspondant à une des millions de variables présentes dans le modèle.

La procédure de prévision d'ensemble commence par l'extraction d'un échantillon fini de valeurs de la distribution décrivant l'incertitude de l'état initial de l'atmosphère. Par exemple, dans un espace des phases à deux dimensions, la distribution de probabilité de l'état initial peut être représentée par un nuage de points autour de la valeur moyenne, où la densité décroît avec la distance par rapport à la moyenne. On peut donc imaginer l'obtention de l'échantillon en effectuant un tirage aléatoire de quelques membres du nuage de points autour de l'état atmosphérique moyen. Ces points sont alors appelés « ensemble des conditions initiales » et chacun d'eux représente un état initial plausible de l'atmosphère compatible avec l'incertitude découlant des observations et analyses.

La démarche la plus simple pour construire une méthode permettant de générer l'ensemble des conditions initiales est de considérer la valeur représentant l'état le plus probable de l'atmosphère (obtenue après analyse des observations) comme étant la moyenne de la distribution de probabilité représentant l'incertitude de l'état initial de l'atmosphère. Les variations autour de cet état moyen peuvent être facilement générées (grâce notamment aux caractéristiques sur l'incertitude des observations), par une méthode de Monte Carlo (en partant d'une même analyse que l'on perturbe aléatoirement, on peut réaliser un certain nombre de prévisions et essayer d'en déduire une formulation probabiliste du temps qu'il fera). En perturbant légèrement l'état initial, c'est-à-dire en le modifiant de façon aléatoire dans les limites que l'on sait être l'amplitude de l'incertitude initiale, on produit une nouvelle prévision qui diffère plus ou moins de la prévision de référence. En reproduisant cette opération un grand nombre de fois, on obtient un ensemble de prévisions qui représente théoriquement tous les états futurs possibles de l'atmosphère. Cependant en pratique, cette méthode fournit des ensembles dont les membres sont trop similaires les uns des autres. La variabilité de l'ensemble des prévisions résultantes paraît sous-estimer l'incertitude de la prévision. Un autre inconvénient de ces méthodes de perturbations aléatoires réside dans la quantité considérable de calculs qu'il faut faire chaque fois que l'on modifie le programme de prévisions.

D'autres méthodes ont été développées ne consistant pas à perturber la situation initiale de façon aléatoire, mais à chercher les champs (il s'agit des zones mais aussi les variables et les intervalles de temps) les plus instables, ceux pour lesquelles une faible erreur d'analyse risque de conduire à une forte erreur de prévisions. Retenons seulement que sont sélectionnées celles des modifications qui conduisent à un écart maximal entre prévision perturbée et prévision de référence après un temps d'intégration donné.

Deux méthodes :
- méthode dite de « breeding » (les perturbations initiales sont obtenues de façon itérative à partir d'une modification arbitraire, en réinjectant dans le modèle les modifications qui ont le taux de croissance le plus élevé)
- méthode basée sur les vecteurs singuliers consistant à rechercher les régions de l'atmosphère les plus sensibles (celles où une petite erreur sur l'état initial est susceptible de croître extrêmement rapidement). Les axes de l'instabilité maximum peuvent être calculés à partir des vecteurs singuliers.

2) Utilisations de la prévision d'ensemble :

La prévision permet plusieurs choses :
- peut **fournir une mesure de la prévisibilité de l'atmosphère** (une faible dispersion des différentes prévisions entraîne une bonne confiance en la prévision et donc une bonne prévisibilité, une grande dispersion entraîne une faible prévisibilité)
- peut montrer des alternatives sur le comportement atmosphérique

Une des utilisations de cet ensemble de prévisions est de considérer la moyenne. En effet, les équations gouvernant l'évolution de l'atmosphère et utilisées pour transformer les conditions initiales en prévisions sont des fonctions non linéaires. On a donc la propriété :

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \neq f(\frac{1}{n}\sum_{i=1}^{n} x_i)$$

Le terme de gauche représente la moyenne de l'ensemble de prévisions tandis que celui de droite est l'unique prévision obtenue à partir de la moyenne de l'ensemble des conditions initiales. Cette moyenne de l'ensemble améliorera la prévision seulement s'il n'y a pas de changements de régimes. La prévision la plus probable n'est pas forcément la meilleure.

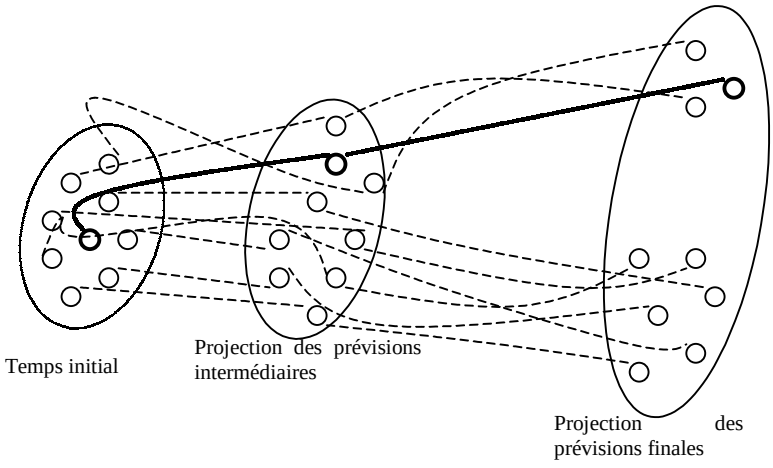


Illustration du concept d'ensemble de prévisions représentée dans un espace des phases à deux dimensions

La ligne en gras représente l'évolution de l'état initial analysé. Les lignes en pointillés représentent l'évolution des autres membres de l'ensemble des conditions initiales. L'ellipse représente la distribution statistique des états atmosphériques initiaux qui sont très proches les uns des autres. Au temps intermédiaire, tous les membres de l'ensemble sont encore raisonnablement similaires. Au temps final, certains membres ont subi un changement de

régime, et représentent différents écoulements de l'atmosphère. La distribution finale est séparée en deux groupes. La dispersion des prévisions étant trop grande, l'état futur de l'atmosphère est très incertain. La moyenne de l'ensemble final de prévision ne représente alors pas l'état probable futur de l'atmosphère.

Un aspect particulier de l'ensemble de prévisions est sa capacité à fournir une information sur la nature de l'incertitude en une prévision. Qualitativement, on aura une grande confiance si la dispersion de l'ensemble est petite et on pourra alors dire que la moyenne est proche de l'état éventuel de l'atmosphère. Par contre, si les membres sont très différents les uns des autres, l'état futur est très incertain.

3) Utilisation en météorologie d'une méthode de classification des prévisions :

L'objectif d'une classification est de synthétiser la grande quantité d'informations disponibles en classant les prévisions. Les méthodes de classement consistent à rassembler les éléments les plus proches les uns des autres pour ne plus les représenter individuellement mais collectivement sous la forme d'éléments types.

(a) Le tubing

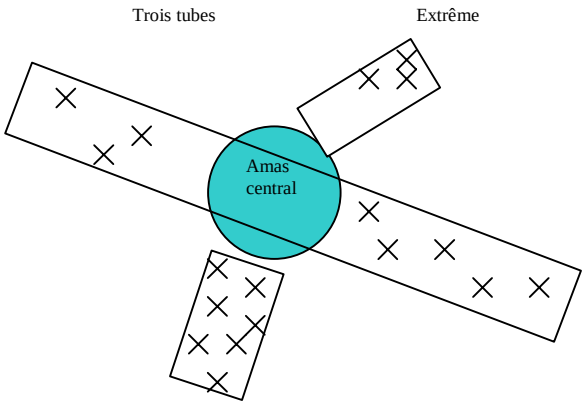
La méthode utilisée en météorologie s'appelle le « tubing ». Elle consiste surtout à mettre en évidence les principales variantes extrêmes qui existent dans la distribution d'ensemble. Elle donne du poids, non seulement au centre de la distribution (moyenne de l'ensemble) mais aussi à ses éléments extérieurs (c'est-à-dire aux prévisions qui diffèrent le plus de la moyenne).

L'hypothèse nécessaire est que la distribution des prévisions est multimodale (la chance de réalisation d'une prévision décroît lorsque l'on s'éloigne de la moyenne).

On délimite un amas central qui regroupe les prévisions les plus proches de la moyenne. La prévision la plus éloignée de l'amas central devient alors l'extrême de ce qu'on appelle le premier tube. Dans ce tube, on regroupe toutes les prévisions dont la distance à l'axe du tube est inférieure au rayon de l'amas central. On réitère le processus jusqu'à ce qu'il n'y ait plus d'extrêmes esseulés. (Une prévision peut appartenir à plusieurs tubes à la fois, et le nombre de tubes dépend de la contrainte imposée sur le rayon de l'amas central).

La moyenne des champs associés à l'amas central est une représentation de la prévision principale la plus probable (elle met en évidence les caractéristiques communes à toutes les prévisions qui ont donc le plus de chances de se réaliser). Les extrêmes de chaque tube représentent des variantes de la prévision la plus probable (les autres prévisions du tube vont dans le « même sens » en étant moins marquées).

Cette méthode de classification fournit donc un état probable et ses variantes. Le seul paramètre d'ajustement est le rayon de l'amas central. Il dépend de l'échéance : plus l'échéance est lointaine plus la dispersion des prévisions est grande et donc plus le rayon doit être grand (on fait aussi en sorte que le nombre de tubes obtenus soit compris entre 1 et 6).



Dans cette représentation schématique du tubing, chaque croix correspond à une des prévisions de la prévision d'ensemble. Si une prévision se résumait à prévoir le centre d'une dépression le schéma serait alors une représentation exacte du tubing. Mais en fait une prévision ne se limite pas à la connaissance de la position du centre d'une dépression. Elle est définie par un ensemble de n valeurs, n étant le produit du nombre de paramètres du modèle par le nombre de niveaux et par le nombre de points de grille.

(b) Elaboration d'un indice de confiance

La météorologie effectue des prévisions jusqu'à sept jours d'échéance assorties d'un indice de confiance. Celui-ci prend la forme d'un chiffre entre un et cinq :

1 très faible confiance
2 faible confiance
3 confiance normale
4 bonne confiance
5 très bonne confiance

Pour définir cet indice de confiance, il faut partir du nombre de tubes correspondant au nombre de variantes de scénarios. En théorie, une confiance très élevée correspond à 0 tube et une confiance très faible à 5 tubes ou plus. Une correction est aussi envisagée lorsque certaines variantes sont considérées comme non significatives. On obtient alors l'indice technique qui qualifie directement le scénario de la moyenne de l'ensemble et non les phénomènes prévus qui lui sont associés (par exemple, on peut être sûr des conditions anticycloniques mais pas de la dissipation des brouillards). Il faut également adapter la valeur objective fournie par l'indice technique avec, pour objectif, de qualifier l'incertitude globale sur le temps. Enfin l'indice de confiance n'est pas indépendant de la précision adoptée pour exprimer la prévision (choix pour présenter une prévision entre un énoncé très précis mais

plus incertain quant à sa vérification future et un énoncé moins précis mais plus sûr quant à sa réalisation). L'indice de confiance de part son caractère subjectif ne correspond donc pas à une grandeur physique mesurable. Il intègre dans une valeur toutes les incertitudes de la prévision (il ne s'agit pas d'une mesure de l'incertitude particulière de chacun des éléments du temps et ne doit pas être confondu avec une probabilité). Il est important de noter qu'il s'agit d'un indice relatif à l'échéance et à la période de l'année. L'indice de confiance permet de connaître le niveau de l'incertitude relativement à l'incertitude normale à laquelle on est en droit de s'attendre pour l'échéance et la période de l'année considérée.

Attention, il convient de rappeler que la prévision de l'incertitude n'est pas une prévision de la fiabilité de la prévision. C'est seulement en moyenne que les prévisions les plus incertaines sont le moins fiables, et c'est cela qu'il s'agit de vérifier pour contrôler la qualité d'un indice de confiance.

(c) Rôle du prévisionniste

Le tubing est une méthode de classification qui permet d'avoir accès à la prévision déterministe la plus probable et de l'assortir au travers d'une estimation de la dispersion des différentes prévisions, d'un indice de confiance qui reste subjectif puisqu'il est établi par le prévisionniste*. Le prévisionniste joue un rôle clé dans la prévision du temps. En premier lieu, il interprète l'ensemble des données d'observations afin de suivre l'évolution de l'atmosphère. Ensuite, au sein de la masse d'informations considérables (satellites, observations supplémentaires non disponibles au moment où a été lancée la prévision…) il doit dégager les éléments essentiels qui détermineront le temps des heures et jours à venir. Il existe diverses causes d'imperfections des modèles numériques comme les lacunes du système d'observations et des méthodes d'analyse, la maille de discrétisation horizontale trop lâche, la représentation incomplète des différents processus physiques. Il est nécessaire de garder une attitude critique lors de l'interprétation des sorties de modèles et de tenter de pallier leurs insuffisances. Le prévisionniste se concentre sur les résultats des modèles numériques et doit valider ou retoucher les prévisions issues de ces modèles. Fort de sa connaissance du climat régional et des limites des modèles, il ajuste, voire modifie, les résultats de la simulation et les traduit en termes de temps observable, comme la durée et l'intensité des précipitations, les températures minimale et maximale du jour, la possible occurrence de brouillards, d'orages ou de rafales de vent. Cette expertise humaine est indispensable pour obtenir une prévision correcte du temps, surtout pour les échéances les plus brèves (de quelques heures à un ou deux jours).

L'analyse critique des observations s'effectue par des méthodes automatiques (comparaison à l'ébauche, statistiques du champ considéré, régularité des observations, historique…). En revanche, le prévisionniste joue un rôle important dans l'analyse météorologique (i.e. quel temps va-t-il faire) des résultats numériques produits par le modèle.

Le prévisionniste a intérêt à disposer du plus grand nombre de prévisions numériques alternatives : à défaut de pouvoir prévoir la réalité avec certitude, il est préférable de connaître un ensemble de réalités vraisemblables, en espérant pouvoir déceler parmi elles certaines convergences ou divergences. C'est l'idée toute simple qui est à la base de la prévision d'ensemble, dont le but est de fournir aux prévisionnistes un nombre aussi grand que possible d'évolutions différentes de l'atmosphère visant à représenter l'ensemble des évolutions qui peuvent découler de l'état initial considéré.

Un système de prévision d'ensemble est destiné à fournir un ensemble de N prévisions de l'état météorologique comme N réalisations indépendantes d'une distribution de

* : le prévisionniste a un rôle similaire à celui de l'exploitant en trafic

probabilité prédite. La prévision d'ensemble permet d'obtenir par exemple, à partir des 50 analyses perturbées, 50 prévisions plus la prévision basée sur l'état initial original dite *prévision de contrôle.*

IV. Vers la prévision probabiliste

La prévision d'ensemble permet de prévoir, pour une échéance donnée, à travers la distribution des diverses valeurs prévues les probabilités associées. Une approche de la prévision d'ensemble consiste à voir dans le résultat de la prévision d'ensemble les distributions de probabilité des différentes variables météorologiques traitées par le modèle numérique. Tout « événement » météorologique peut se ramener à une combinaison d'occurrences élémentaires, dont la probabilité est connue grâce à la distribution fournie par l'ensemble. Ainsi, la probabilité de l'événement « beau temps chaud » définie par « pas de précipitations, température supérieure à 25°C et vent inférieur à 5 m/s », est donnée immédiatement par le nombre de prévisions pour lesquelles ces différentes occurrences sont prévues simultanément, rapporté au nombre total de prévisions.

Un système de prévision probabiliste ne permet pas de prédire l'état de l'objet physique considéré mais prédit la loi de probabilité de cet état (fonction de distribution de probabilité pdf). En pratique il prédit un ensemble fini de valeurs numériques qui en paramétrant définissent la pdf et peuvent être considérées comme des réalisations indépendantes de la pdf (obtention d'un échantillon).
Un tel système pourra produire des probabilités pour l'occurrence d'événements spécifiques.
C'est donc une fonction de probabilité qui doit être calculée en chacun des points et pour chacune des grandeurs atmosphériques.

1) Système du pauvre : moyen de construction d'un ensemble de prévisions

Nous avons que l'une des possibilités pour construire un ensemble de prévisions était de faire varier les conditions initiales, dans les limites que l'on sait (il existe plusieurs manières pour changer les données d'entrée). Ces méthodes sont valables puisque le système d'équations physiques est sensible à des variations des données d'entrée. Néanmoins, il existe une autre solution pour l'obtention de l'ensemble final. Cette solution appelé « système de pauvre » ne demande pas de variations des conditions initiales.

En considérant une variable météorologique associée à un modèle de prévision, cette construction, pour une échéance donnée, ne requiert pas d'intégrations successives du modèle à partir de conditions initiales différentes. Elle utilise les prévisions déterministes issues du même modèle et calculées sur un intervalle de temps antérieur au temps de la prévision.

On suppose une variable scalaire x(t) associée à un modèle de prévision.
Hypothèse : on possède un échantillon de K prévisions antérieures provenant du même modèle $\{x_c(t_k), k \in [1\ldots K]\}$ ainsi que les observations correspondantes (ce qu'on appelle les valeurs vérifiées) $\{x_v(t_k), k \in [1\ldots K]\}$.
A l'instant t seule la prévision contrôlée est disponible $x_c(t)$.

Méthode simple de construire un ensemble de prévisions avec N membres :

Extraire du passé (à savoir de l'échantillon $\{x_c(t_k), k \in [1\ldots K]\}$) les N valeurs les plus proches de $x_c(t)$ et prendre comme prévisions les observations correspondantes (attention : N<<K).

En notant I l'ensemble des N indices compris entre 1 et K correspondant aux données ainsi retenues l'ensemble de prévision de x(t) est $\{x_v(t_k), k \in I\}$ (auquel on rajoute la prévision contrôlée au temps t $x_c(t)$).
Cet ensemble peut être manipulé comme n'importe quel ensemble de prévisions.

Une des principales difficultés de la prévision probabiliste est l'évaluation de la qualité d'un tel système. Dans le paragraphe suivant, après avoir défini les deux principales approches sur lesquelles est basée cette mesure de la qualité, seront présentés les différents coefficients de mesure pour, dans un premier temps des événements individuels (pluie, neige..), puis pour des variables directement issues du modèle de prévisions.

2) Evaluation de la qualité d'un système de prévision probabiliste :

Il existe une corrélation entre la distribution de l'ensemble des prévisions et l'erreur des prévisions observées à posteriori. Un système de prévision probabiliste ne permet pas de prédire l'état de l'objet physique considéré mais prédit la loi de probabilité de cet état (fonction de distribution de probabilité pdf). En pratique il prédit un ensemble fini de valeurs numériques qui en paramétrant définissent la pdf et peuvent être considérées comme des réalisations indépendantes de la pdf (obtention d'un échantillon).

Le problème majeur pour mesurer la qualité de prévisions d'un tel système est que l'objet prédit* et l'objet observé sont de deux natures différentes. Il devient alors impossible d'estimer une « distance » entre l'objet prédit et l'objet observé. En effet la fiabilité (qui indique dans quelle mesure la prévision est conforme à la réalité observée) d'une prévision déterministe s'apprécie généralement par des mesures directes de l'écart entre valeurs prévues et valeurs observées (par exemple, une erreur moyenne de température) ou bien par des indicateurs de la performance d'une prévision d'occurrence (ex : taux de non détection d'un brouillard).

Les diverses mesures pour évaluer cette qualité sont basées sur différentes approches :

- Evaluer l'accord entre les probabilités prédites et les observations⇒ *robustesse* du système c'est-à-dire accord entre les distributions des probabilités prédites à priori et les fréquences des observations à posteriori. Une prévision fiable est avant tout une prévision « calibrée » (correspondance entre probabilité prévue et fréquence observée).
- Analyser la variabilité dans les distributions des probabilités prédites (*résolution*)

La mesure de la qualité d'un système de prévisions doit représenter au mieux ces deux propriétés.

- *Prévisions probabilistes pour des événements individuels :*

Pour obtenir la courbe de robustesse des prévisions statistiques d'un événement E donné, on trace la fréquence des observations de l'événement E p'(p) en fonction de la probabilité prédite p (robustesse si p'(p)=p). Une courbe proche de la diagonale p'= p signifie une forte robustesse.
(Remarque : bien que la taille de l'échantillon soit grande, chaque probabilité p n'est pas prédite assez souvent pour que la valeur correspondant à la fréquence de l'événement p'(p) soit stable.)

Il est possible de définir un score de « capacité » qui mesure l'habileté à obtenir un meilleur score que celui obtenu par la prévision de référence. La capacité d'un système de

*: par prédit nous entendons la sortie de la modélisation et non pas la prévision élaborée par les prévisionnistes

prévisions d'ensemble est estimée en utilisant comme référence une unique prévision de contrôle de l'ensemble.

Définition du Brier Score : $B = \frac{1}{M}\sum_{i=1}^{M}(p_i - o_i)^2$ où M est le nombre de réalisations du système (*une prévision est définie par un ensemble de M valeurs, M étant le produit du nombre de paramètres du modèle par le nombre de niveaux et par le nombre de points de grille. C'est un point d'un espace à M dimensions*). Pour chaque réalisation i, p_i est la probabilité prédite et o_i prend la valeur 1 ou 0 selon que l'événement E est observé ou non.
Le parfait système déterministe correspond à B=0.

Représentation continue de B :

$B = \int_0^1 [(1-p')p^2 + p'(1-p)^2]g(p)dp$ où g est la fréquence avec laquelle p est prédite.

Décomposition du score B :

$B=B_c+ B_v$

$B_c = \int_0^1 (p'-p)^2 g(p)dp$ qui mesure la robustesse ou la consistance du système (=0 pour p'(p)=p)

$B_v = \int_0^1 p'(1-p')g(p)dp = -\int_0^1 (p'-p_c)^2 g(p)dp + p_c(1-p_c)$ où $p_c = \int_0^1 p'(p)g(p)dp$

p_c représente la référence climatologique.
B_v mesure la dispersion autour de p_c des fréquences observées à posteriori.

Normalisation de B :

$BSS = 1 - \frac{B}{p_c(1-p_c)}$ score qui croît de 0 à 1 (système parfait)

L'exactitude avec laquelle la qualité du système est évaluée est inévitablement limitée (erreur due à l'observation et au nombre fini de réalisations).
La taille de l'ensemble de prévisions notée N (par exemple, N=50 c'est-à-dire on fait 50 prévisions au lieu d'une) a son importance ; l'impact numérique d'une augmentation de N est plus grand lorsque les probabilités prédites ont une petite dispersion que lorsqu'elles ont une grande dispersion.

Autre graphique intéressant, *la courbe de ROC* (Relative Operating Characteristic) : courbe basée sur la stratification des observations. Elle dépend uniquement de la dispersion des probabilités p'(p). La résolution d'un système de prévisions probabilistes pour l'arrivée d'un événement E est mesurée par l'aire sous la courbe de ROC.

- *Prévisions probabilistes pour des variables individuelles :*

Variable scalaire x(t)

La pdf pour x est définie par N valeurs x_i rangées dans l'ordre croissant (ce qui donne N+1 intervalles).
L'observation correspondante x_v (si elle provient de la réalisation de la pdf) doit tomber avec une fréquence 1/(N+1) dans chaque intervalle.
L'histogramme de la position de x_v en fonction des intervalles donnés par les x_i donne une mesure de la consistance du système (un système parfait correspond à un histogramme plat).
s_j est la population dans chaque intervalle.

$\Delta = \sum_{j=1}^{N+1} (s_j - \frac{M}{N+1})^2$. Si x_v est distribué uniformément alors $\Delta = \frac{MN}{N+1}$.

Une autre mesure de la robustesse statistique peut être obtenue grâce à la quantité :
D=ENSK-ENSP
Où ENSK=(x_v –m)² (m est la moyenne de la pdf prédite) et ENSP est la variance de la pdf.

Pour un grand nombre de réalisations du système il est possible de travailler sur les valeurs moyennes.
$\overline{D} = \overline{ENSK} - \overline{ENSP}$
Cette quantité doit tendre vers zéro.

Remarque sur les limites de la prévision probabiliste :

La prévision d'ensemble est un outil mathématique permettant d'élaborer de véritables prévisions probabilistes sous forme de probabilité d'occurrence d'un événement. Certains paramètres du temps peuvent être aisément prévus sous forme probabiliste par la prévision d'ensemble (paramètres directement prévus par le modèle). Par exemple le gel pouvant être considéré comme une température négative à 2 m du sol, la prévision d'ensemble peut fournir automatiquement une prévision du type « risque de gel à 70% ». Par contre pour des phénomènes tels que la neige ou l'orage dont l'occurrence est liée à la combinaison de plusieurs paramètres, ils deviennent plus complexes à évaluer sous forme probabiliste.

V. Assimilation de données

Reste que pour utiliser tout ce qui vient d'être présenté il faut néanmoins avoir recours au modèle numérique de prévision et donc l'initialiser. Or la méconnaissance de l'état initial n'a pas été résolue. En effet comme il l'a été souligné en première partie, les observations ne peuvent décrire sans ambiguïté l'état de l'atmosphère et ne sont pas forcément relevées aux points de grille (points pour lesquels on calcule l'écoulement de l'atmosphère).

Des méthodes d'assimilation de données sont aujourd'hui de plus en plus développées pour palier à ces problèmes.

L'assimilation de données est la combinaison d'un modèle numérique de l'écoulement et d'observations distribuées dans le temps. L'objectif recherché est d'obtenir la meilleure prévision pour un instant donné et de trouver la meilleure méthode à priori pour corriger cette prévision à l'aide des observations.

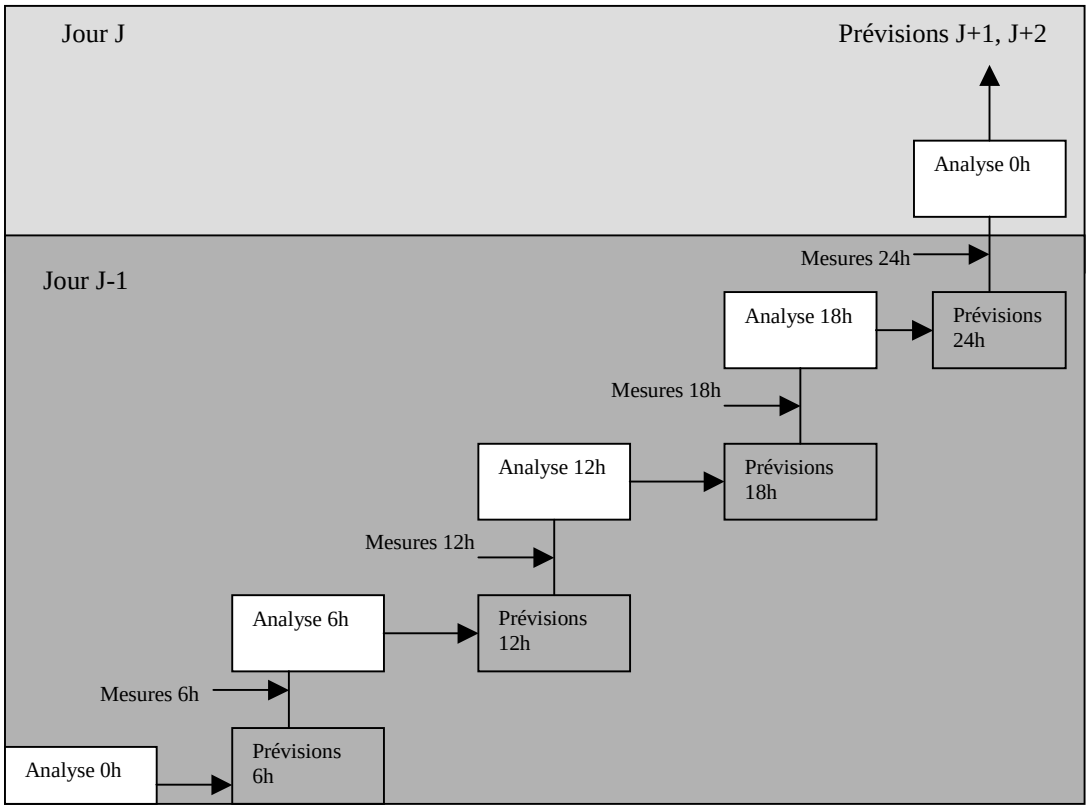
Le problème à résoudre se ramène donc à trouver la meilleure méthode pour corriger cette prévision à l'aide des observations.

On constate que les modèles de prévisions numériques tels qu'ils sont présentés utilisent des propriétés physiques plus que des propriétés statistiques. La modélisation de l'écoulement de l'atmosphère est en effet basée sur des lois physiques vérifiées par les paramètres caractérisant l'évolution de l'état de l'atmosphère. La statistique intervient alors pour faire des prévisions locales et non pour prédire l'écoulement de l'atmosphère. Néanmoins en ce qui concerne l'analyse objective des observations les propriétés statistiques vont être utilisées pour déterminer l'état initial le plus probable de l'atmosphère.

L'analyse objective a tout d'abord été réalisée à l'aide de méthodes d'interpolation géométriques n'utilisant aucune propriété statistique, puis avec des méthodes de corrections successives d'une ébauche fournie par un modèle de prévision. **La prise en compte des propriétés statistiques des champs de variables météorologiques**, fondement des méthodes d'interpolation optimale, a été une étape importante permettant de tenir compte des caractéristiques propres des diverses observations disponibles et de tirer avantage des liens existants entre les champs à analyser.

Actuellement, on ne peut traiter par assimilation que des problèmes linéaires ou linéarisables. Les phénomènes non-linéaires ne peuvent pas encore être pris correctement en compte dans le modèle d'assimilation. C'est le cas notamment du cycle de l'eau atmosphérique et c'est la raison pour laquelle les images satellitaires ne sont pas assimilées. Des progrès sont en cours, notamment par les approches « incrémentales » (i.e. on représente un processus non-linéaire par une série –que l'on espère convergente- de processus linéaires).

Dans ce paragraphe, après avoir représenté schématiquement un cycle d'assimilation de données, seront présentés d'un point de vue théorique les fondements des méthodes d'interpolation optimale. Dans le deuxième paragraphe, une présentation générale de la théorie de l'estimation en assimilation de données sera exposée. Enfin, le principe de l'assimilation variationnelle quadridimensionnelle permettant de prendre en compte l'évolution temporelle des paramètres sera développée succinctement.



Pour faire l'analyse des champs météorologiques pour le jour J à 0 heure, on repart de l'analyse du jour précédent à 0 h. On commence par faire une prévision à 6h d'échéance, qui est utilisée avec toutes les mesures parvenues entre temps, pour effectuer une analyse à 6h. On recommence cette opération avec les mesures de 12, 18 et 24 h pour obtenir l'analyse à 0h du jour J qui sert à faire les prévisions.

Ce cycle d'assimilation correspond à l'analyse objective.

1) Théorie des méthodes d'interpolation statistiques :

L'expérience passée sur le comportement de l'atmosphère est utilisée comme source d'informations pour la détermination des poids de l'interpolation.
Si on note f_{it} la valeur réelle numérique d'une variable météo au point i à l'instant t.
Pour les observations

$f_{it}^{obs} = f_{it} + \Delta f_{it}$ où Δf_{it} est l'erreur due à l'observation (aux observations est associée une erreur)

On définit des coefficients qui considèrent la déviation entre la vraie valeur et la valeur obtenue par le modèle de prévision (considérée comme une valeur préliminaire du champ, par exemple peut aussi être une valeur moyenne climatologique).

$m_{ij} = \overline{(f_{it} - f_{it}^P)(f_{jt} - f_{jt}^P)}$ covariance pour une variable f entre 2 points i et j

$\mu_{ij} = \frac{m_{ij}}{\sqrt{m_{ii} m_{jj}}}$ autocorrélation

$M_{f_i g_j} = \overline{(f_{it} - f_{it}^P)(g_{jt} - g_{jt}^P)}$ cross covariance entre 2 variables

$\mu_{f_i g_j} = \frac{M_{f_i g_j}}{\sqrt{Var_i Var_j}}$ crosscorrélation

On utilise n observations f_{it}^{obs} (i=1…n) dans le voisinage du point de grille g pour calculer, pour un certain temps t, la valeur analysée f_{gt}^{NA}. Cette valeur est calculée comme une combinaison linéaire de la valeur préliminaire (obtenue par le modèle de prévision) et des déviations $f_{it}^{obs} - f_{it}^P$.

$$f_{gt}^{NA} = f_{gt}^P + \sum_{i=1}^{n} p_i (f_{it}^{obs} - f_{it}^P)$$

Les p_i (poids d'interpolation)sont obtenus en minimisant $E = \overline{(f_{gt} - f_{gt}^{NA})^2}$.
Les poids doivent donc vérifier

$$\sum_{i=1}^{n} (m_{ik} + d_{ik}) p_i = m_{kg} \quad k = 1...n$$

avec $m_{ij} = \overline{(f_{it} - f_{it}^P)(f_{jt} - f_{jt}^P)}$ covariance des erreurs des prévisions

et $d_{ij} = \overline{(f_{it}^{obs} - f_{it})(f_{jt}^{obs} - f_{jt})}$ covariance des erreurs des observations

Cette méthode d'interpolation statistique a été suivie par une méthode de corrections successives. L'idée de base de cette méthode est de corriger le champ préliminaire (champ prévu par le modèle pour tous les points de grille) itérativement pendant l'horizon de plusieurs analyses.. Les prévisions sont effectuées pendant que les observations sont utilisées pour produire l'analyse de l'état du système. A l'issue d'une prévision on utilise les résultats de l'analyse pour corriger les résultats de la prévision et ce nouvel état corrigé représente alors le point de départ pour la nouvelle prévision.

2) Présentation générale de la théorie de l'estimation :

- Cas scalaire : système décrit par une variable d'état scalaire x inconnue

x_o observation d'erreur associée ε_o .
x_m état prévu par le modèle d'erreur associée ε_m .
Les erreurs sont des variables aléatoires gaussiennes indépendantes d'écart types σ_o et σ_m connus.
La forme variationnelle du problème revient à minimiser la somme des écarts de l'état du système aux observations et à la prévision du modèle (somme des carrés pondérés par les variances des erreurs).

$$J(x) = \frac{1}{2}[\sigma_o^2 (x - x_o)^2 + \sigma_m^2 (x - x_m)^2]$$

Le minimum est obtenu pour

$$x = \frac{\sigma_m^2}{\sigma_o^2 + \sigma_m^2} x_o + \frac{\sigma_o^2}{\sigma_o^2 + \sigma_m^2} x_m$$

- Cas vectoriel : l'atmosphère est décrite par un grand nombre de variables définies en de nombreux points du temps et de l'espace.

On suppose que les observations ont lieu au même instant.

Dans le cas général, les observations ne sont pas dans le même espace que les variables du modèle et de l'analyse (par exemple, pour la prévision de la température, le satellite recueillant les observations permet la seule observation des rayonnements localisés sur sa trace), d'où l'utilisation d'un opérateur C en général non linéaire mais linéarisable au voisinage du point considéré. On supposera donc C linéaire.

On obtient la formulation variationnelle en minimisant la somme, pondérée par les matrices de variance covariance d'erreur, des écarts entre l'état analysé du système et les observations d'un côté, les prévisions du modèle de l'autre.

$$J(x) = (x - x_m)^t M^{-1} (x - x_m) + (Cx - x_o)^t O^{-1} (Cx - x_o)$$

Minimum obtenu pour $x_a = x_m + MC^t (CMC^t + O)^{-1} (x_o - Cx_m)$.

$x_o - Cx_m$ représente la différence entre les valeurs observées et les valeurs que l'on devrait avoir si l'atmosphère était exactement décrite par l'état du modèle (vecteur résidu ou innovation).

La dimension du problème est celle de x, c'est-à-dire le nombre de points de la grille de discrétisation.

Les formules présentées ci-dessus présentent la forme la plus générale de l'interpolation optimale. Cette méthode est communément utilisée avec quelques simplifications car d'une part, la dimension des matrices à inverser étant de l'ordre de 10^4 ou 10^5 ,une résolution exacte est trop coûteuse, et d'autre part la détermination des matrices de covariance des erreurs est difficile.

L'analyse objective possède deux inconvénients. En théorie, elle ne permet que l'utilisation des mesures effectuées de façon synchrone avec l'heure d'analyse (comme l'analyse est effectuée toutes les 6h et que les stations météorologiques diffusent généralement les mesures au sol toutes les 3h, la moitié de ces mesures n'est pas utilisée). De plus on se sert de toutes les mesures effectuées dans les 24h qui précèdent l'analyse mais pas de celles qui suivent. Avant de réaliser un cycle d'assimilation à partir des analyses du jour J pour obtenir une ébauche de l'analyse au jour J+1 à 0h on peut penser qu'il devrait être possible d'utiliser les données arrivées dans cette période pour améliorer l'analyse de départ et non pas seulement l'analyse d'arrivée.

Ces méthodes d'analyse objective permettent donc de trouver la meilleure estimation possible de l'atmosphère à l'**instant donné** compte tenu d'une prévision obtenue à l'aide d'un modèle et d'observations de l'atmosphère **à cet instant**. Or les observations dont on dispose sont réparties dans le temps. Il faut donc effectuer plusieurs analyses séquentiellement.

Une des possibilités pour effectuer une analyse séquentielle est l'utilisation du filtre de Kalman

Présentation succincte du filtre de Kalman en assimilation de données :

Notations :

- L'espace des états du modèle (**E**) est l'ensemble des états possibles du modèle et a pour dimension le nombre de ses variables independantes.

Les variables dans **E** sont notées x.

- L'espace des observations (**F**) est l'ensemble des observations possibles et a pour dimension le nombre de mesures independantes.

Les variables dans **F** sont notées y.

- Le forecast (x^f) est l'état du modèle *à-priori*, 'est à dire avant l'assimilation, il est assorti d'une incertitude représentée par sa matrice de covariance d'erreur : **P** (La matrice de covariance d'erreur *à-posteriori* est notée $\mathbf{P}^f$).
- Les observations (y^o) sont les données mesurées, elle sont aussi assorties d'une incertitude : **R**.

L'incrément ajouté lors de l'assimilation peut être un élement quelconque de **E**, mais, pour limiter les calculs, on le restreint souvent à un sous-espace de **E** : l'espace de contrôle.

- L'espace de contrôle (**C**) est celui dans lequel est calculé la correction effectuée lors de l'assimilation, c'est un sous-espace de **E**.

Les variables dans **C** sont notées z.

Dans toute la suite, nous négligerons cet aspect, et nous supposerons que **E** = **C.**
L'opérateur d'observation (H) fournit le jeu d'observations correspondant à un état dans **E**.
Le modèle (M) simule l'avance temporelle du système physique. La matrice de covariance d'erreur du modèle lorsqu'on passe du temps t_1 au temps t_2 est $\mathbf{Q}_{(t2,t1)}$.

Equations de base :

Le maximum de vraisemblance correspond au minimum de la forme quadratique :
$2\mathbf{J}(x,y) = [x-x^f]\mathbf{P}^{-1}[x-x^f] + [y-y^O]\mathbf{R}^{-1}[y-y^O]$
Sous la contrainte $y=H(x)$

Ce problème classique peut être résolu suivant une stratégie d'approche linéaire :

On suppose ***H*** linéaire, ce qui ramène notre problème à la résolution du système linéaire suivant :
$[x^a-x^f] = (\mathbf{P}^{-1}+\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}[y^O-H(x)] = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T+\mathbf{R})^{-1}[y^O-H(x)]$
La matrice de covariance d'erreur **à posteriori** est :
$\mathbf{P}^{-1}_{\text{assimilée}} = \mathbf{P}^{-1}_{\text{forecast}} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$

La première stratégie consiste à effectuer tous ces calculs, inversions matricielles comprises : c'est le Filtre de Kalman (il nécessite donc des calculs lourds dès que la dimension de **E** devient élevée).

Le principal défaut de cette méthode est qu'une observation donnée influe seulement sur l'évolution future de l'état mais n'est pas utilisée pour corriger les états passés à cause du caractère en une seule « passe » de l'assimilation. Elle permet donc l'obtention d'une bonne représentation de l'état à la fin d'une période d'assimilation mais pas sur tout l'intervalle en question.

3) Assimilation variationnelle quadridimensionnelle :

L'approche « assimilation variationnelle » des données d'observation permet de prendre en compte l'information fournie par une grande variété de systèmes. Afin de minimiser les erreurs des champs analysés des différentes variables météorologiques, ceux-ci doivent être ajustés les uns par rapport aux autres (c'est-à-dire que n'importe quelle relation entre les variables est forcée d'être complètement ou partiellement réalisée).

On suppose que le modèle décrivant l'atmosphère est parfait (c'est-à-dire les erreurs dues aux prévisions sont nulles soit d'après les notations précédentes M=0).
L'évolution du système est gouvernée par $x_{k+1} = Fx_k$ (1).
La formulation variationnelle du problème conduit à considérer
$J(x) = \sum_{k=0}^{N} (Cx_k - x_{o,k})^t O^{-1} (Cx_k - x_{o,k})$ (2) où x est l'ensemble des états du modèle aux instants successifs k=0…N liés entre eux par l'équation d'évolution du modèle.
Minimiser la distance J sous la contrainte (1) produit pour tout temps k le meilleur estimateur linéaire sans biais de l'état du système compte tenu de toutes les informations disponibles.

Cette méthode revient à un problème de minimisation de la somme des carrés entre le modèle et les observations pondérée par l'inverse de la matrice variance covariance sous la contrainte que x est solution d'une équation d'évolution différentielle. Dans ce cas, on a recours au Lagrangien, or en pratique des problèmes de résolution existent étant donné le grand nombre de paramètres et la complexité des équations. Il devient alors obligatoire de transformer le problème initial avec contrainte en une séquence de problème de minimisation sans contrainte : trouver un état initial tel que la solution correspondante du modèle d'évolution minimise la fonction mesurant la distance aux observations. Cette transformation et la résolution de ce type de problèmes demandent l'utilisation de l'adjoint non développée dans ce rapport.

En conclusion, cette méthode de résolution met surtout en évidence la difficulté de la mise en place de l'assimilation de données. Le principe d'assimilation est un concept relativement simple à comprendre par contre sa mise en œuvre de part la présence d'équations non linéaires et du grand nombre de paramètres, reste extrêmement compliquée et demande des résolutions mathématiques complexes ainsi qu'une puissance de calculs importante.

Deuxième partie :
Application des concepts d'évaluation utilisés en météorologie aux prévisions de trafic

La première partie avait pour objectif la présentation de la philosophie des méthodologies utilisées dans le cas d'une prévision du temps. Dans cette discipline, l'incertitude portant sur les données initiales ainsi que l'aspect chaotique de l'atmosphère ont obligé la recherche d'un moyen pour prévoir l'incertitude sur l'état futur. A défaut de pouvoir prévoir la réalité avec certitude, il est préférable de connaître un ensemble de réalités vraisemblables, en espérant pouvoir déceler parmi elles certaines convergences ou divergences. C'est l'idée toute simple qui est à la base de la prévision d'ensemble, dont le but est de fournir aux prévisionnistes un nombre aussi grand que possible d'évolutions différentes de l'atmosphère visant à représenter l'ensemble des évolutions qui peuvent découler de l'état initial considéré.

L'analogie recherchée entre la météorologie et le trafic réside dans les approches de calcul de l'incertitude des prévisions. Nous pensons que les prévisions de trafic (à l'horizon variant d'un an à quelques jours) peuvent se faire selon différentes conditions de l'offre routière (l'état de la route, la disponibilité des autres modes de transports…) et la demande de déplacements.

L'objectif de cette partie sera donc de mettre en application les concepts de prévision d'ensemble au trafic routier. A partir du modèle du dispositif Bison Futé (BF) qui permet d'obtenir une prévision du débit journalier un an à l'avance, on va chercher à construire un ensemble de prévisions avec le système du pauvre. L'objectif de la prévision d'ensemble sera de donner à priori une validité dans la prévision. Actuellement on ne peut avoir qu'une idée de la qualité de la prévision à posteriori (i.e. en comparant la prévision à la valeur réelle connue une fois l'année passée). Une autre application de cet ensemble peut aussi mettre en évidence dans quelle mesure la prévision est mauvaise, c'est-à-dire savoir s'il y a, à priori, une sous-estimation ou une sur-estimation. La construction d'un indice de confiance, comme en météorologie, de la qualité de la prévision est une possibilité.

Après avoir présenté le modèle de prévisions utilisé dans le cadre du dispositif de Bison Futé basé sur un modèle GLM, ainsi que les résultats obtenus sur un capteur donné, nous appliquerons la prévision d'ensemble. On rappellera alors la méthodologie de construction de l'ensemble puis ses différentes utilisations. Les résultats appliqués au capteur de St Arnoult seront proposés pour différents types de jours.

I. Présentation générale du modèle du dispositif Bison Futé :

Bison Futé a été conçu comme un dispositif de gestion des départs en congés. Cette conception a conduit à privilégier deux axes de travail :

- La prévision pour mieux anticiper sur les encombrements aux points critiques
- L'information aux usagers comme moyen d'agir sur les comportements.

L'objectif est donc de prévoir à l'horizon d'un an le trafic journalier puis de diffuser ses prévisions afin d'obtenir un changement de comportement des usagers en vue d'un meilleur écoulement du trafic. Pour créer ce dispositif, il apparaît logique de faire appel à des techniques de modélisation mathématique afin de prévoir le trafic journalier et horaire. C'est ainsi que, dès la fin des années 80, un logiciel « Bison Futé » voit le jour pour lequel différentes méthodes de prévisions sont élaborées (Couton, Danech-Pajouh, Debeauvais, 1996). Prévues à l'origine pour le calcul des prévisions du trafic de la période du 15 Juin au 15 Septembre, celles-ci ont depuis été étendues, dans un premier temps à l'ensemble des périodes de départ et de retour de vacances et de ponts, puis à l'année entière.

L'historique utilisé par le logiciel couvre les années allant de 1987 à 1997, il est composé des résultats de 80 stations SIREDO. Les mesures de débit de chaque station sont

relevées toute l'année. L'intervalle de mesure est d'une heure, ce qui permet de disposer de 8 760 valeurs chaque année, dans chaque sens de la circulation. La répartition des stations utilisées par le logiciel n'est pas homogène sur le territoire français. Certaines zones sont privilégiées pour leur importance dans les flux nationaux et européens.

1) Présentation de la méthodologie :

Nous appelons **débit journalier**, que nous assimilons au terme trafic, le nombre de véhicules passés en un point donné au cours d'une journée.

Le modèle de prévision du trafic journalier étudié est une version améliorée de celui élaboré conjointement par le SETRA et le CETE Nord Picardie (voir Ziani et Danech, nov. 1998). Il est actuellement utilisé dans le cadre du dispositif Bison Futé pour l'élaboration du calendrier des jours à forte circulation : calendrier qui paraît un an à l'avance, ce qui explique l'horizon de prévision du modèle. A aussi long terme, il est donc impossible de tenir compte du comportement dynamique de la circulation, c'est à dire de l'écoulement du flux de trafic, ainsi que des événements exceptionnels tels que les grèves, les intempéries ou certains travaux.

Seules les caractéristiques calendaires du jour considéré (type de jour, fêtes, ponts…) restent donc disponibles pour la modélisation, d'où l'utilisation d'une approche de type analyse de la variance utilisant des variables explicatives qualitatives issues de ces événements calendaires. Mais comme le montre les figures 1.1 et 1.2 qui sont des représentations sur le capteur St Arnoult, la seule modélisation calendaire ne permet pas d'expliquer la tendance annuelle du trafic. On applique donc le modèle linéaire au rapport des débits sur le Trafic Moyen Journalier Annuel (TMJA), et non aux débits bruts. La méthode de prévision donne donc aux débits une structure multiplicative,

$$q(j,m,a) = TMJA(a) \times q_r(j,m,a) \qquad (1)$$

où un jour est caractérisé par sa date j/m/a, avec j, le jour, m le mois et a l'année.
TMJA(a) désigne le Trafic Moyen Journalier Annuel de l'année a, q(j, m, a) les débits journaliers réels et q_r(j, m, a) les débits journaliers relatifs pour le jour de date j/m/a.

Quant au processus de calcul des prévisions, il suit le schéma suivant, c'est-à-dire une prévision en deux temps :

- Calcul des Trafics Moyens Journaliers Annuels de l'historique : TMJA(a)
(a désigne l'année considérée)
- Prévision du Trafic Moyen Journalier Annuel de l'année b de prévision :
$TMJA^*$(b)
- Calcul des débits relatifs de l'historique : q_r(j, m, a) tels que,

$$q_r(j,m,a) = \frac{q(j,m,a)}{TMJA(a)} \qquad (2)$$

- Prévision des débits journaliers relatifs du jour de date j/m/b : q_r^*(j, m, b)
- Calcul des débits prédits : q^*(j, m, b) tels que,

$$q^*(j,m,b) = q_r^*(j,m,b) \times TMJA^*(b) \qquad (3)$$

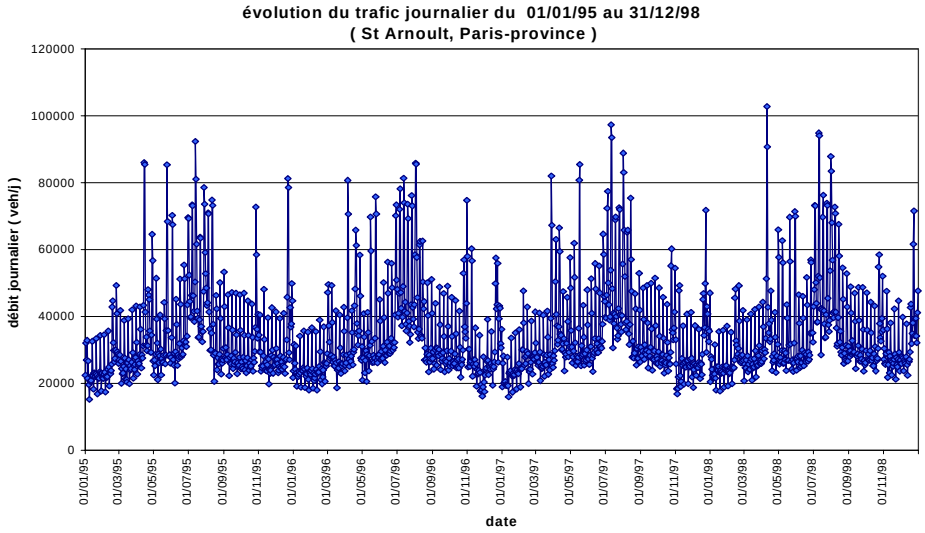


Figure 1.1

2) Prévision du Trafic Moyen Journalier Annuel (TMJA) :

Celle-ci fait appel aux techniques des séries chronologiques qui extrapolent l'évolution des TMJA à partir de la connaissance des mesures de l'historique. Pour une meilleure qualité de la prévision, c'est à partir du **Trafic Moyen Journalier Mensuel** (TMJM) que la prévision est effectuée : ceux-ci sont évalués, puis prédits à l'aide d'une **méthode de type Winters** qui permet de prendre en compte à la fois la tendance globale de l'évolution du trafic, et la périodicité annuelle de celui-ci (voir Figure 1.2). Le calcul et la prévision des TMJA sont ensuite obtenus en réalisant les moyennes des douze TMJM de l'année considérée, pondérées par le nombre de jours de chaque mois. D'où un processus de calcul divisé en deux étapes :

- Calcul et prévision des Trafics Moyens Journaliers Mensuels (TMJM) :

TMJM(i, a) où $1 \le i \le 12$ désigne le mois, et a l'année considérés dans l'historique correspondant aux cinq dernières années. L'historique est réduit par rapport à l'ensemble des dix années de données disponibles afin de garantir une meilleure qualité des prévisions.

- Le TMJA(a) est alors la moyenne pondérée des TMJM(i, a) de l'année a.

$$TMJA(a) = \frac{1}{\sum_{i=1}^{12} n_i} \sum_{i=1}^{12} n_i \; TMJM(i, a) \qquad (4)$$

où n_i est le nombre de jours du $i^{ème}$ mois de l'année a.

De même, pour le calcul du TMJA prédit pour l'année b : $TMJA^*(i, b)$, $1 \le i \le 12$, on obtient,

$$TMJA^*(b) = \frac{1}{\sum_{i=1}^{12} n_i} \sum_{i=1}^{12} n_i \; TMJM^*(i, b) \qquad (5)$$

où n_i désigne cette fois-ci le nombre de jours du $i^{ème}$ mois de l'année b.

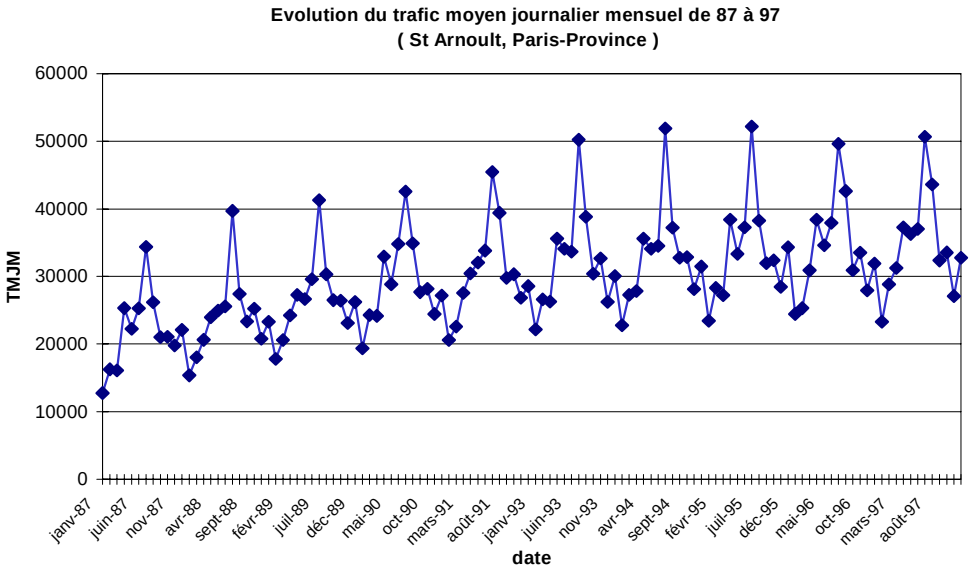


Figure 1.2

La méthode de Winters employée pour prédire les TMJM est à tendance linéaire, et de périodicité 12.

Nous présentons succinctement la méthodologie de calcul.

Notons $\text{tmjm}(t) = \text{TMJM}(i, a)$ (6)

avec, $t = i + 12(a - a_1)$ (7)

où $1 \le i \le 12$ désigne le numéro du mois considéré, et $a_1 \le a \le a_5$, représente l'année (nous rappelons que l'historique utilisé est de 5 ans).

Les TMJM sont alors modélisés par la formule,

$$\text{tmjm}(t) = [\beta_0(t) + \beta_1(t)t] \times SN_t(t) + \varepsilon_t \quad (8)$$

où les termes $\beta_0(t)$, $\beta_1(t)$ et $SN_t(t)$ sont des fonctions susceptibles d'évoluer lentement avec le temps : $[\beta_0(t) + \beta_1(t)t]$ caractérise la tendance et $SN_t(t)$ le facteur saisonnier de périodicité 12. Les TMJM sont alors estimés par des formules du type,

$$\hat{\text{tmjm}}(t) = [b_0(t) + b_1(t)\, t] \times sn_t(t) \quad (9)$$

où $b_0(t)$, $b_1(t)$ et $sn(t)$ sont définies par récurrence en fonction de l'instant précédent et du facteur saisonnier correspondant (voir le livre de Bowerman et O'Connell, p.403-408, 1993) .

Pour $T = 12(a_5 - a_1 + 1)$, la dernière donnée de l'historique pour les TMJM, les prévisions effectuées à partir de cet instant sont alors de la forme,

$$\mathrm{tmjm}^{*}(T+\tau)=\left[b_0(T)+b_1(T).\tau\right]\mathrm{sn}_{T+\tau-12}(T+\tau-12) \qquad (10)$$

où $\mathrm{sn}_{T+\tau-12}(T+\tau-12)$ désigne l'estimateur du facteur saisonnier de l'année précédente, et si une prévision doit être faite pour un horizon dépassant 12 mois, les estimateurs utilisés pour les facteurs saisonniers seront les derniers calculés pour la période de l'année correspondante. Par exemple, pour le $i^{ème}$ mois de l'année de prévision, $b=a_5+1$ qui suit directement la dernière année de l'historique,

$$\mathrm{TMJM}^{*}(i,b)=\left[b_0(T)+b_1(T).i\right]\mathrm{sn}_{T+i-12}(T+i-12) \qquad (11)$$

et pour la prévision du TMJM du $i^{ème}$ mois de l'année b+1,

$$\mathrm{TMJM}^{*}(i,b+1)=\left[b_0(T)+b_1(T).(i+12)\right]\mathrm{sn}_{T+i-12}(T+i-12) \qquad (12)$$

En pratique, l'erreur relative d'estimation d'un TMJA dépasse rarement les 3%.

3) Prévision des débits relatifs :

Les débits relatifs $q_r(j,m,a)$ sont définis par :

$$q_r(j,m,a)=\frac{q(j,m,a)}{TMJA(a)} \qquad (13)$$

où $q(j,m,a)$ désigne le débit journalier du jour de date j/m/a, et $\mathrm{TMJA}(a)$ le Trafic Moyen Journalier Annuel de l'année a**.**

Ces débits relatifs sont estimés, puis prédits, à l'aide **d'un modèle de régression de type analyse de la variance dont les variables explicatives sont caractéristiques des événements calendaires** (type de jour, ponts, départs en vacances, retours...). Ces variables étant de type qualitatif, elles ne peuvent être utilisées sous leur forme d'origine. Chacune est donc transformée en autant de variables quantitatives qu'elle possède de modalités, ces nouvelles variables ne prenant que deux valeurs numériques : 1 si la modalité concernée est vérifiée, et 0 sinon. C'est finalement une relation linéaire entre ces différentes variables quantitatives qui est cherchée pour expliquer le débit relatif, relation qui est extrapolée pour prédire les données futures.

Notons que **les observations des trois dernières années de l'historique ont une pondération deux fois plus importante que celle des autres observations, ce qui permet de prendre en compte l'évolution significative du comportement des automobilistes au cours des dernières années**.

Le modèle utilisé est de la forme classique :

$$q_r = X\beta + \varepsilon \qquad \varepsilon \approx \aleph_n\left(0, \sigma^2 I_n\right) \qquad (14)$$

où

- $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ est un vecteur de paramètres que l'on cherche à estimer,
- $q_r = (q_{r_1}, q_{r_2}, \ldots, q_{r_n})'$ est le vecteur des n observations de l'historique pour les débits relatifs.
- $X = (x_i^j)_{\substack{1 \le j \le n \\ 1 \le i \le p}} = (\chi_1 \chi_2 \ldots \chi_p)$ est la matrice $(n \times p)$ des n observations de l'historique des p variables explicatives sous forme disjonctive.

La matrice X devient :

$$X = \left(\begin{array}{c|c|c|c|c} 1 & & & & \\ \vdots & X_1 & X_2 & \cdots & X_q \\ 1 & & & & \end{array} \right)$$

où les X_i, $1 \le i \le q$ sont les matrices blocs correspondant aux tableaux disjonctifs $(n \times p_i)$ associés aux n observations de chacune des q variables qualitatives : chaque variable explicative i possède un nombre p_i de modalités ($p_i \ge 2$) et est transformée en p_i variables numériques quantitatives correspondant chacune à l'indicatrice de l'une des modalités. Ainsi, $p = 1 + \sum_{i=1}^{q} p_i$ est le nombre de colonnes de la matrice X.

- $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ est le vecteur $(n \times 1)$ des erreurs d'estimation de chacune des n observations. Il est supposé distribué selon une loi normale de moyenne nulle et d'écart type σ.

Nous obtenons comme forme d'équation pour le modèle de prévision,

$$q_r^* = X^* \beta \qquad (15)$$

où X^* représente la matrice $(n^* \times p)$ des p variables explicatives des n^* données à prédire.

β est alors estimé par la méthode **des moindres carrés**, ce qui donne l'estimateur $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p\right)$ minimisant la somme des carrés des résidus, soit $\hat{\beta}$ tel que,

$$\left\| q_r - X\hat{\beta} \right\|^2 = \underset{\beta \in \Re^p}{Min} \left\| q_r - X\beta \right\|^2 = \underset{\beta \in \Re^p}{Min} \left(q_r - X\beta\right)' \left(q_r - X\beta\right) \qquad (16)$$

Notons que ceci revient également, par dérivation par rapport à β, à résoudre le système linéaire,

$$(X'X)\beta = X'q_r \qquad (17)$$

C'est d'ailleurs la résolution de ce système qui permet d'évaluer numériquement β.

β n'est pas estimable puisque (X'X) n'est pas de plein rang, l'équation (17) admet au moins une solution mais elle n'est pas unique. Par contre, une fonction linéaire Cβ peut être estimable. On a donc la propriété :

sous l'hypothèse $\mathrm{Ker(X)} \subset \mathrm{Ker(X^*)}$ $\hat{q}_r^*$ est unique.

Une fois obtenu $\hat{\beta}$ le meilleur estimateur de β qui permet de trouver la meilleure adéquation du modèle par rapport aux données de l'historique, on extrapole la formule pour obtenir les prévisions. La prévision du débit relatif du jour de date j/m/b, $q_r^*(j,m,b)$, peut alors être facilement obtenue en multipliant le vecteur de ses variables explicatives $x^*(j,m,b)$, par le paramètre $\hat{\beta}$, ce qui donne comme formule,

$$\hat{q}_r^*(j,m,b) = x^*(j,m,b)\hat{\beta} \qquad (18)$$

Ce principe, généralisé à l'ensemble de l'horizon de prévision, donne alors le vecteur (n*×1) des estimations des prévisions,

$$\boxed{\hat{q}_r^* = X^* \hat{\beta}} \qquad (19)$$

En suivant le modèle proposé par le SETRA et le CETE de Nord-Picardie, nous avons choisi d'utiliser des paramètres explicatifs correspondant, soit à des informations qualitatives simples comme le type de jour, le mois ou la présence d'un départ en vacances, soit à des informations résultant d'interactions d'ordre deux entre ces informations simples, comme le type de jour croisé avec la présence d'un départ en vacances. Le détail de ces variables est donné ci-dessous, avec le codage leur correspondant.

Typj: type de jour (lundi, mardi, ...)
Mois: le mois avec éventuellement plusieurs modalités pour certains mois.
zone(1,2,3): vacances scolaires par zone. Si premier jour un lundi ou dernier jour un vendredi, le samedi et le dimanche sont inclus dans les vacances. Correspond au codage du calendrier officiel.
Dep(1,2,3): spécifie les trois premiers jours des vacances scolaires par zone On code les veilles, 1er, 2ème, et 3ème jours de départ.
Pon: typologie des ponts en 7 classes selon la position du jour férié dans la semaine.

Jour férié	Jours codés(autour du jour férié)
Dimanche	vendredi à lundi
Lundi	vendredi à lundi
Mardi	vendredi à mercredi
Mercredi	mardi à jeudi
Jeudi	mercredi à lundi

Vendredi	jeudi à lundi
Samedi	vendredi à lundi

Pont2: pont du vendredi à lundi lorsque le lundi est férié. Précise si le lundi est inclu dans les vacances scolaires. Tous les jours du pont : du vendredi au lundi, sont codés.
Ren(1,2,3): rentrée scolaire par zone. (dernier jour de vacances, ainsi que sa veille et son avant-veille, premier jour de rentrée)
Typep: numérotation des jours d'un pont dans l'ordre croissant.
Typef: type de fête (tous les jours du pont sont codés)
Fet: une modalité par fête. Soit 11 au total a priori, en fait 9 (le jour de l'an, l'Assomption et le 11 Novembre sont codés de façon identique). Tous les jours du pont sont codés.
Dist: deux ponts dans la même quinzaine. Tous les jours du pont sont codés.
Sem: numéro de la semaine dans l'année. Même numérotation que le calendrier officiel.

On a finalement le modèle qui explique le débit relatif par une fonction linéaire des variables explicatives :

mois, fet, typef, typj, pon, pont2, dep1, dep2, dep3, sem, zone1, zone2, zone3, dist×pon, mois×sem, mois×typj, mois×´dep1, mois×dep2, mois×dep3, typj×pon, typj×dep1, typj×dep2, typj×dep3, typj×zone1, typj×zone2, typj×zone3, typj×ren1, typj×ren2, typj×ren3, typep×fet, typef×fet, typj×fet, typep×zone1, typep×zone2, typep×zone3.

4) Application aux données relatives au péage de St Arnoult :

Le travail effectué porte sur les données du capteur de St Arnoult dans le sens de circulation de Paris vers la province. L'historique est composé des données de Janvier 1987 à Décembre 1997 (hormis l'année 92 supprimée pour cause de mauvaise qualité des observations). Pour l'ensemble de l'année 1998, on cherche à prédire les débits journaliers.

La version actuelle du modèle s'appuie sur les procédures du logiciel SAS : utilisation de la procédure FORECAST pour la prévision des TMJM, et de GLM pour celle des débits relatifs.

Un filtrage des données de l'historique a également été réalisé afin d'éliminer les données aberrantes : dans un premier temps tous les jours de débits journaliers trop faibles (inférieurs à 1000 véhicules) sont ôtés, puis dans un deuxième temps, après un premier calcul de prévisions, on supprime les jours dont l'erreur relative absolue d'estimation est supérieure à 30%. Ces données aberrantes se rencontrent lorsque le capteur tombe en panne, ou lorsque des événements « exogènes » exceptionnels, comme des grèves, des accidents, ou des travaux, viennent perturber l'écoulement normal du trafic. Pour le capteur de St Arnoult, environ 40 jours de mesures sur les 3650 données initiales de l'historique sont ainsi ôtées.

Pour pallier au problème de non unicité de $\hat{\beta}$ pour la modélisation des débits relatifs, la procédure GLM de SAS fixe arbitrairement à 0 un nombre de coordonnées de $\hat{\beta}$ égal à $p - \text{rang}(X'X)$. Le calcul des coordonnées restantes revient ainsi à résoudre un système linéaire dont la solution est unique. Notons également que la procédure GLM de SAS ne

donne pas de prévision si une nouvelle modalité d'une variable explicative apparaît pour un jour de l'horizon (pas d'estimation pour le $\hat{\beta}_i$ correspondant, on est dans le cadre où la prévision n'est pas estimable) : il est alors nécessaire d'éliminer les variables explicatives susceptibles d'être à l'origine d'un tel phénomène pour obtenir des prévisions pour tout l'horizon de prévision. C'est ce qui a déjà été appelé la procédure de ventilation des variables. En pratique, l'algorithme utilisé est donc le suivant :

- Filtrage des données de trafic : on élimine les jours dont les débits journaliers sont aberrants (trop faibles car inférieurs à 1000 véhicules par jour)
- Calcul des TMJM par année pour l'historique.
- Prévision des TMJM avec la procédure FORECAST.
- Calcul des TMJA réels et prédits comme moyenne pondérée des TMJM.
- Calcul des débits relatifs de l'historique.
- Estimation du trafic relatif par la procédure GLM
- Calcul des résidus pour l'estimation des débits relatifs.
    - Filtrage des données de l'historique : on élimine tous les jours dont l'erreur relative absolue d'estimation est supérieure à 30% (ce seuil est un choix arbitraire qui ne repose sur aucune règle statistique)
- Ventilation des variables explicatives.
- Deuxième prévision du trafic relatif par GLM (c'est elle qui est retenue).
- Prévision du trafic comme produit du trafic relatif prédit et du TMJA prédit.

(a) Présentation générale des résultats obtenus pour tout type de jours

Pour confirmer l'efficacité du modèle présenté dans le paragraphe précédent, il a donc été testé sur le capteur de St Arnoult. A partir des données des années 1987 à 1997, l'objectif est de prédire le trafic journalier pour l'année 98. L'étude des résidus présentée ci-dessus permet de valider ou non le modèle.

Soit $\hat{q}(j,m,a)$ le débit estimé du jour dont la date est j/m/a. De même, pour b l'année de prévision (b=1998), on note $\hat{q}^*(j,m,b)$ le débit prédit du jour j/m/b. Nous considérons que l'erreur de prévision sur les TMJA est négligeable.

Pour bien analyser les résidus, il est nécessaire de séparer l'étude de ceux de l'**estimation** de celle de ceux de la **prévision**. En effet, d'une façon évidente, les premiers sont moins importants et mieux distribués car ils résultent directement de l'erreur d'ajustement du modèle aux données réelles, alors qu'il y a dégradation des seconds, le comportement du trafic des jours à venir n'étant pas entièrement déductible de celui des jours passés.

On appelle alors résidu de l'estimation du jour de date j/m/a,

$$res(j,m,a) = [\,\hat{q}(j,m,a) - q(j,m,a)\,] \qquad (25)$$

et résidu de la prévision du jour j/m/b,

$$res^{*}(j,m,b) = \left[\hat{q}^{*}(j,m,b) - q(j,m,b) \right] \tag{26}$$

Pour évaluer l'ajustement du modèle, nous avons calculé deux types d'erreurs globales :

- l'*erreur quadratique relative moyenne* **err1**

Pour l'estimation,

$$err1 = \sqrt{\frac{\sum_{j,m,a} [\hat{q}(j,m,a) - q(j,m,a)]^2}{\sum_{j,m,a} q(j,m,a)^2}} \tag{27}$$

Pour la prévision,

$$err1 = \sqrt{\frac{\sum_{j,m,b} \left[\hat{q}^{*}(j,m,b) - q(j,m,b)\right]^2}{\sum_{j,m,b} q(j,m,b)^2}} \tag{28}$$

- l'*erreur quadratique moyenne* **err2**. Cette erreur a l'avantage de donner une indication sur l'ordre de grandeur de l'erreur par rapport au trafic moyen.

Pour l'estimation,

$$err2 = \sqrt{\frac{\sum_{j,m,a} [\hat{q}(j,m,a) - q(j,m,a)]^2}{n}} \tag{29}$$

Pour la prévision,

$$err2 = \sqrt{\frac{\sum_{j,m,b} \left[\hat{q}^{*}(j,m,b) - q(j,m,b)\right]^2}{n^{*}}} \tag{30}$$

Le *tableau 1.1* présente les résultats obtenus avec les différentes erreurs. Nous avons également ajouté l'étude des histogrammes des résidus, soit $res(j,m,a)$ pour l'estimation, et $res^{*}(j,m,b)$ pour la prévision (*Figures 1.3* et *1.4*)

	err1 (%)	**err2 (nbre véhicules)**	**Débit réel moyen**
Estimation	7,49	2463	30128
Prévision	11,55	4314	34704

Tableau 1.1

La qualité des prévisions est donc acceptable, et de plus, les résidus res (j,m,a) semblent normalement distribués.

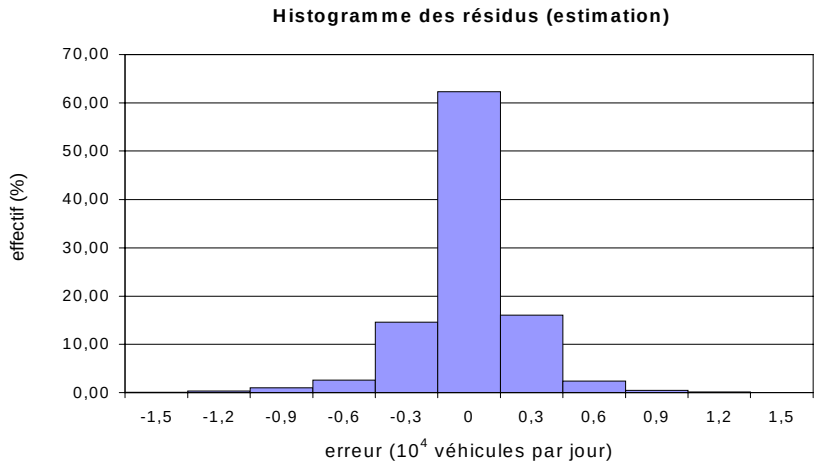


Figure 1.3

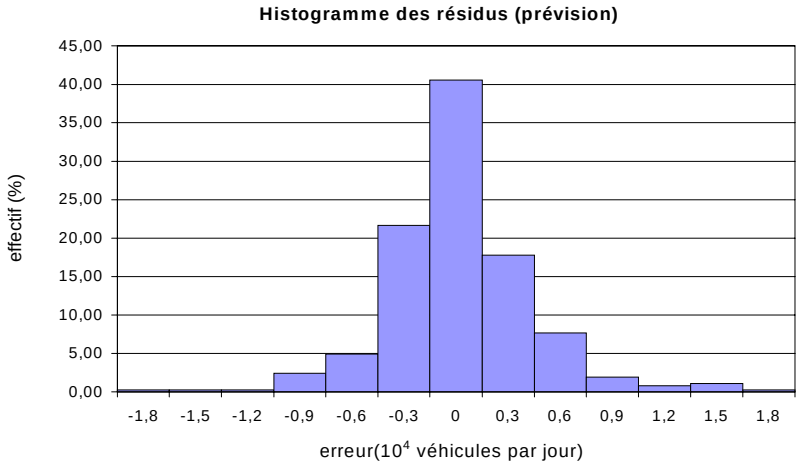


Figure 1.4

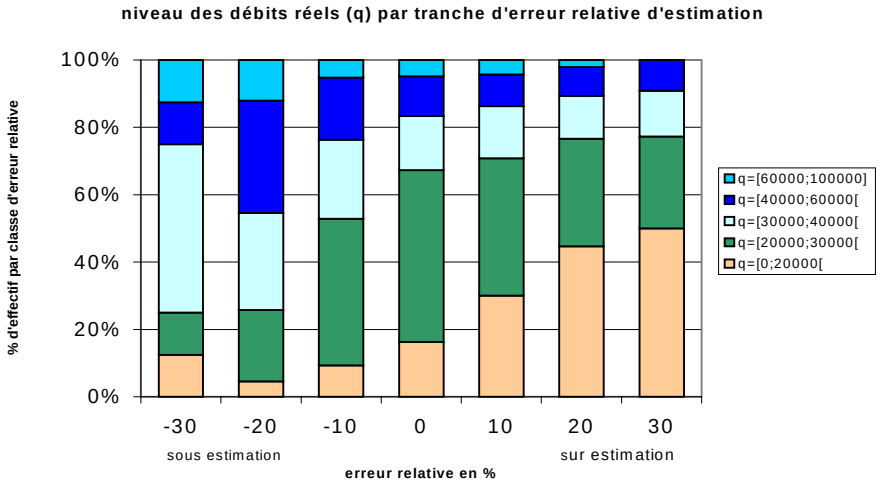


Figure 1.9

La *Figure 1.9* révèle, quant à elle, une répartition inégale des résidus relatifs de l'estimation $\mathrm{res}(j,m,a)/q(j,m,a)$, en fonction du niveau de débit réel : les faibles débits tendent à être surestimés alors que le contraire se produit pour les débits les plus importants. Ce constat a d'ailleurs une influence plus significative pour les prévisions des jours à forts débits, puisqu'une erreur relative importante entraîne une erreur absolue d'autant plus grande que le niveau de débit du jour considéré l'est.
Cette disparité des erreurs relatives en fonction du niveau du débit nous suggère d'aller plus loin dans l'étude des résidus. Pour cela, nous avons décidé d'étudier en détail les résultats obtenus pour les jours « exceptionnels », c'est à dire ceux possédant des caractéristiques calendaires exceptionnelles : pont, départ ou retour de congés scolaires.

(b) Etude des jours exceptionnels

Ces jours « exceptionnels » représentent environ le tiers des données. Le figure 1.10 donne une idée de la répartition des débits réels permettant ainsi de voir que le débit moyen se situe aux alentours de 25000 véhicules (à voir).

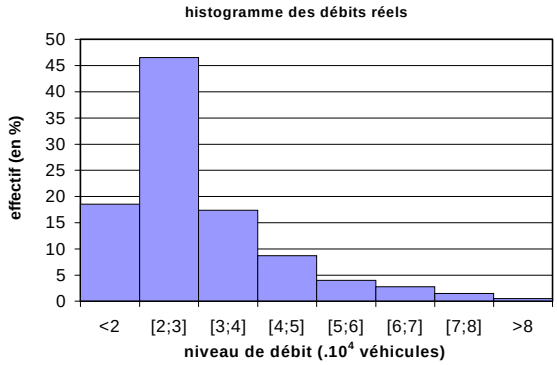


Figure 1.10

La figure 1.11 montre clairement que ces jours où les caractéristiques calendaires sont non ordinaires concernent essentiellement les jours dont le trafic journalier est important.

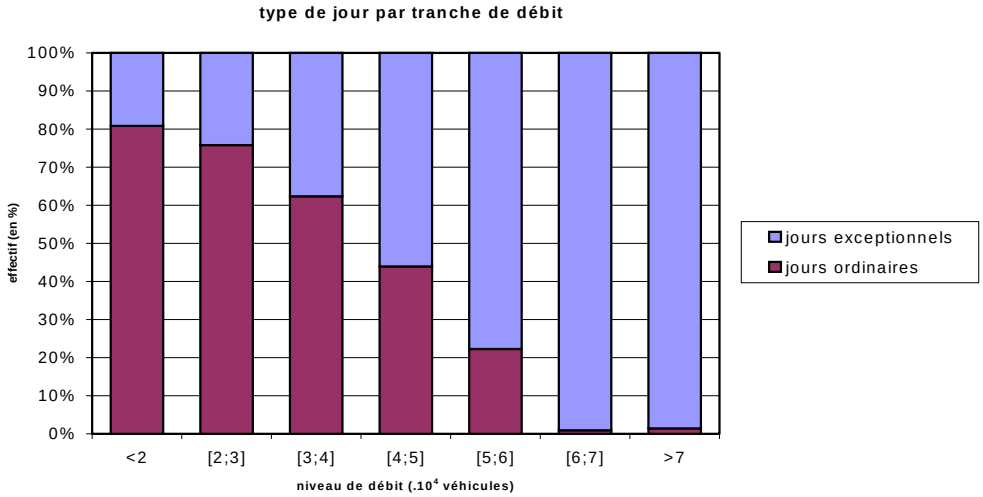


Figure 1.11

C'est donc la qualité des prévisions de trafic pour ces jours qui est la plus intéressante. Nous avons donc réalisé, à partir des résidus utilisés dans la section 1.4.1, une étude spécifique pour ces jours exceptionnels : calcul de leurs erreurs quadratiques propres, de leur moyenne des résidus bruts (voir *Tableau 1.2*), tracé de leur histogramme des résidus pour la prévision et pour l'estimation (*Figures 1.13* et *1.14*). Nous y avons également ajouté une étude de la proportion de jours « exceptionnels » par tranche d'erreur relative d'estimation (voir *Figure 1.12*).

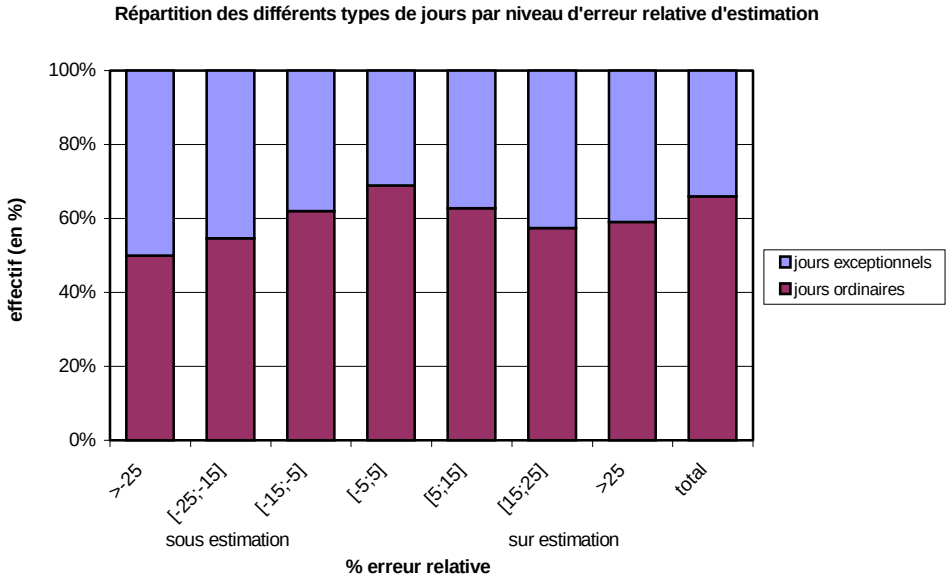


Figure 1.12

	err1 (%)	err2 (nbre véhicules)	Débit réel moyen
Estimation	7,83	3252	37910
Prévision	11,80	5385	41779

Tableau 1.2

Ces résultats montrent une dégradation de la qualité de l'adéquation du modèle aux données réelles par rapport à celle obtenue pour l'ensemble des données (voir *Tableau 1.2*). De plus, parmi les observations dont les résidus relatifs sont importants, nous observons une proportion de jours « exceptionnels » plus importante que la moyenne, et ce notamment s'il y a sous estimation (voir *Figure 1.12*). Ceci est d'ailleurs confirmé par la comparaison des histogrammes des résidus des prévisions avec ceux obtenus pour l'ensemble des données (voir *Figures 1.3*, *1.4*, *1.13* et *1.14*) : la proportion de données pour lesquelles on a des erreurs importantes est plus grande.

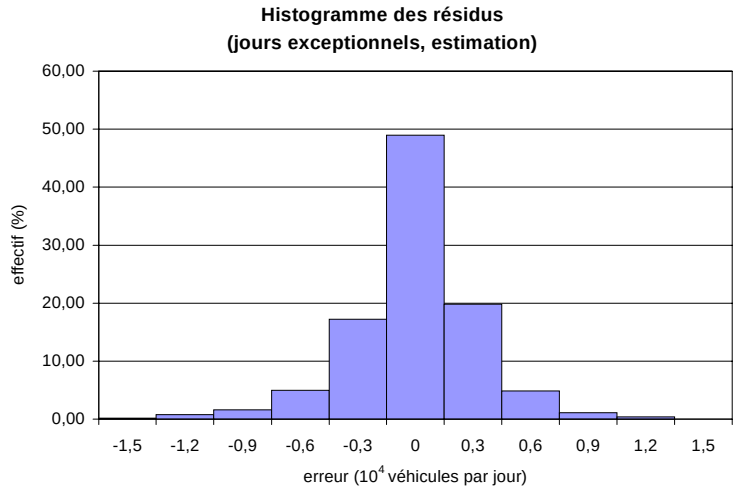


Figure 1.13

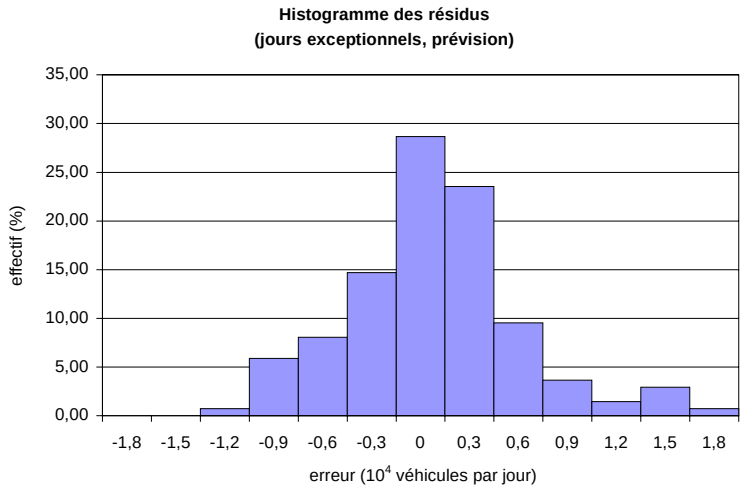


Figure 1.14

Cette étude montre que, si les prévisions sont dans l'ensemble acceptables, elles sont en revanche d'une qualité inégale selon le type de jour considéré. Les débits des jours « ordinaires » sont en général bien estimés et bien prédits, alors qu'il n'en est pas toujours de même pour les jours « exceptionnels ». C'est, en effet souvent pour ce type de jour que les débits sont les plus mal estimés, alors qu'ils sont souvent supérieurs à la moyenne. Cette baisse de la qualité des prévisions est donc un facteur qui nuit à celle du modèle, car des jours à fort débit non prévus peuvent être une source de non détection d'encombrements et donc un handicap pour la mise en place du dispositif Bison Futé, et l'intérêt des informations communiquées aux automobilistes.
Il faut cependant souligner que ces résultats représentent un réel progrès par rapport à ceux obtenus avec les anciens modèles utilisés par le logiciel Bison Futé (voir Couton, Danech-Pajouh et Debeauvais, 1996, p.20-21). Ceux-ci, en effet, obtenaient des prévisions grâce à un principe d'analogie calendaire, trop peu précis pour donner des résultats vraiment satisfaisants.

Le principe d'évaluation du modèle, tel que nous l'avons vu dans le paragraphe précédent, reste une procédure qui ne peut être faite qu'à posteriori analysant l'erreur commise entre la prévision et la valeur réelle. Un des enjeux des méthodes de prévision d'ensemble utilisées en météorologie (voir première partie) est qu'elles essaient de prévoir l'incertitude de l'état futur. En effet, l'analogie recherchée entre le trafic et la météorologie réside dans ce calcul d'incertitude. L'intervention de variables exogènes telles que le temps, le comportement de l'automobiliste, une grève…. qui sont imprévisibles à un horizon aussi lointain rend incertain l'état du trafic et par là la prévision. L'objectif du deuxième paragraphe sera donc, après avoir présenté la prévision d'ensemble par le système du pauvre, d'appliquer cette méthodologie dans le cadre du capteur de Saint Arnoult en se concentrant particulièrement sur les jours dits exceptionnels. On regardera la différence entre l'évaluation faite à priori et celle à posteriori en analysant avec plus de précision les jours pour lesquels l'évaluation à posteriori annonce une forte sous estimation.

II. Prévision d'ensemble : système du pauvre

1) Le système du pauvre :

L'objectif de cette partie est la construction d'un ensemble de prévisions afin d'obtenir une validité dans la prévision pour le jour j. Comme en météorologie, pour qualifier la prévision dite de contrôle réalisée avec le modèle, on va créer un ensemble contenant N prévisions pour le même jour et la même variable.

Cette démarche appelée prévision d'ensemble porte exclusivement sur le débit relatif ce qui permet d'enlever toute tendance annuelle (puisque des comparaisons avec le passé vont être effectuées, l'ensemble doit être homogène). En travaillant sur le débit relatif, on enlève donc tout phénomène de tendance.

Ce type d'ensemble va être construit pour tous les jours de l'année de prévision (que ce soit les jours ordinaires ou exceptionnels). Néanmoins une différence est faite selon le jour à prévoir dans la considération de l'historique.

(a) Principe de construction de l'ensemble de prévision :

Cette construction est basée sur le système du pauvre (cf partie 1).

On désire construire un ensemble pour la prévision du débit relatif $\hat{q}_r^*(j, m, b)$ obtenue par le modèle GLM de l'année b (b=1998) du **jour j** du mois m. Pour cela on a, à notre disposition les estimations issues du même modèle de l'historique qui a servi à la modélisation.

La méthode de construction de l'ensemble se fait en deux étapes :

- La première étape consiste à définir des conditions de ressemblance à travers les variables calendaires afin de définir les jours de l'historique les plus semblables à celui de la prévision. Cet ensemble appelé ensemble des jours semblables sera noté S_h.

- La deuxième étape choisit dans S_h ceux dont l'estimation calculée par le modèle est la plus proche de la prévision pour le jour j.

  ✓ *Première étape : Sélection selon un critère qualitatif issu des variables explicatives calendaires*

Le premier travail consiste à sélectionner dans cet historique les jours semblables au jour j de prévision d'un point de vue calendaire. Pour cette sélection, on a donc recours aux variables explicatives du modèle GLM (voir premier paragraphe). Toutefois, toutes ces variables ne peuvent être utilisées. En effet, si trop de contraintes de ressemblance sont imposées, l'historique S_h ne contiendra pas suffisamment de jours semblables voire même aucun. Or l'objectif final est d'obtenir un ensemble contenant suffisamment de prévisions pour avoir une dispersion représentative des phénomènes pouvant se produire.

- Si le jour j est un jour **ordinaire**, les jours semblables sont tous les jours ordinaires des années 87 à 97 du même type (c'est-à-dire si j est un mardi on prend tous les mardis ordinaires des années 87 à 97 sans tenir compte du mois). L'historique comportera entre 250 et 450 jours (selon le type de jours). **Pour les jours ordinaires, seul le type de jours est un critère de sélection**.

- Si le jour j est un jour **exceptionnel**, on tient alors compte de la saison (en effet les déplacements des jours exceptionnels sont différents selon la saison). Un jour exceptionnel correspond à un trafic journalier important et est détecté via des variables calendaires (ex : départ vacances scolaires, ponts…). Pour ces jours, la prévision calculée par le modèle du dispositif BF commet une erreur importante.

  Pour un jour exceptionnel j des mois de juillet, août 1998, l'historique comportera tous les jours exceptionnels de la mi-juin à la mi-septembre des années 87 à 97. Parmi ceux-ci on ne retiendra que ceux dont le type de jours (lundi, mardi…) est similaire au jour de prévision. Cette période correspond à l'été.
  Pour un jour exceptionnel j des mois de septembre à mars 1998, on effectue la même opération que précédemment en prenant dans l'historique les jours de mi-septembre à mars. Cette période correspond à l'hiver.

  Dans ces deux cas concernant les jours exceptionnels, interviennent à la fois le type de jours et le type de saison.

Pour un jour exceptionnel j des mois d'avril et mai 1998, la démarche est un peu différente. Au départ on ne tient pas compte de la saison et on considère dans l'historique tous les jours exceptionnels du même type (par exemple tous les lundis exceptionnels). On effectue le système du pauvre mais pour un ensemble de prévisions comportant plus de jours. Ensuite on supprimera les jours qui d'un point de vue calendaire nous sembleront aberrants.

✓ *Deuxième étape :nouvelle sélection selon un critère quantitatif*

Algorithme :

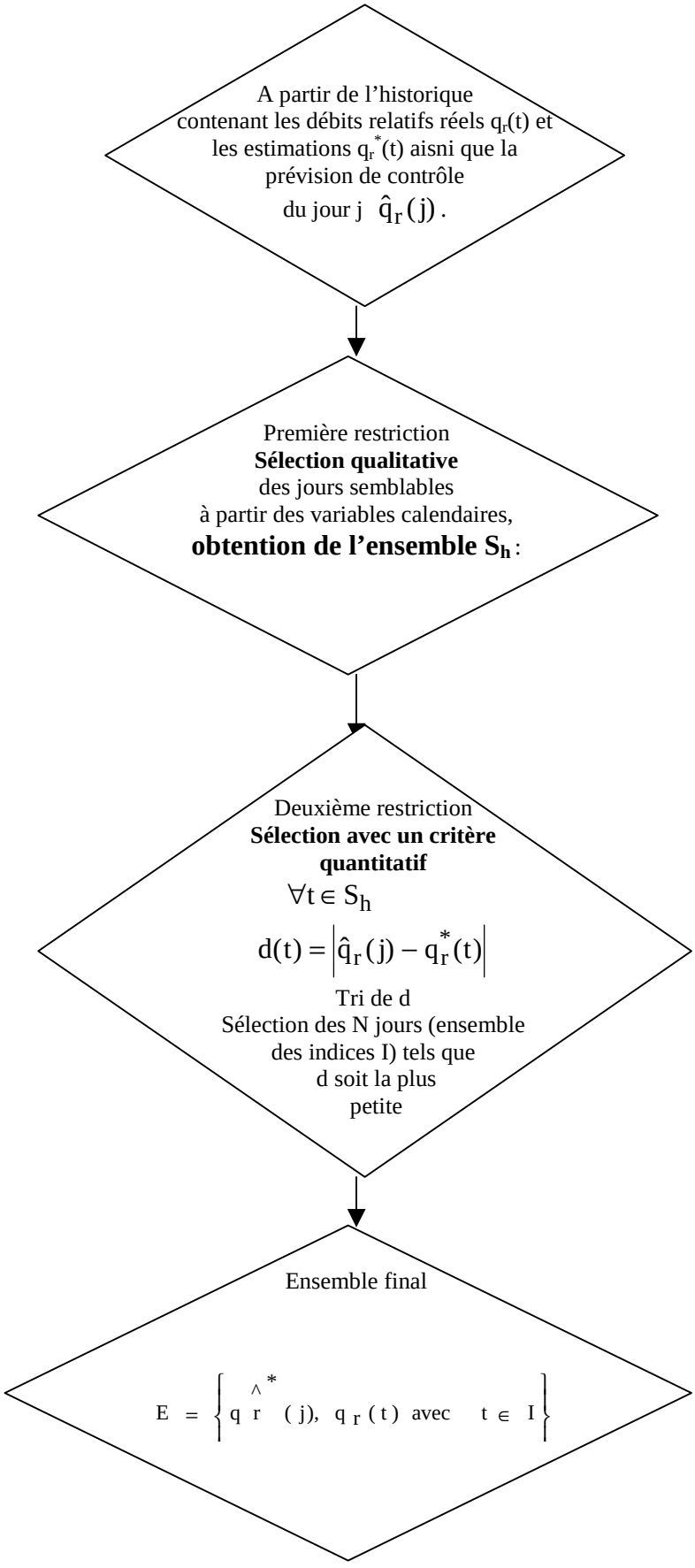
Soit j le jour de l'année 1998 pour lequel on veut construire l'ensemble. Pour ce jour j on connaît le mois m, le type de jours typ_j (lundi , mardi , mercredi…) ainsi que le débit relatif prévu $\hat{q}_r^*(j)$.

- Pour chaque jour semblable de S_h t on connaît la valeur réelle du débit relatif $q_r(t)$ ainsi que l'estimation $q_r^*(t)$ obtenue à partir du modèle GLM qui nous a permis d'obtenir la prévision.
- On calcule la quantité $d(t) = \left| q_r^*(t) - \hat{q}_r^*(j) \right| \quad \forall\, t \in S_h$.
- Parmi les jours de S_h on sélectionne les N jours tels que d(t) est la plus petite possible. (N<<Card (S_h)). Soit I l'ensemble des indices des jours ainsi sélectionnés ($I \subset S_h$).
- L'ensemble de prévisions pour le jour j est alors $E = \left\{ \hat{q}_r^*(j), q_r(t) \text{ avec } t \in I \right\}$ ainsi Card (E)=N+1.

Les estimations de l'historique n'ont servi qu'à choisir les jours se rapprochant le plus du jour j (au sens de la prévision) mais ensuite ne sont considérés dans l'ensemble E final que les **débits réels** de l'historique.

Remarque : la manière d'obtenir l'ensemble final ne tient pas compte des erreurs puisqu'on ne tient compte que des différences entre la prévision et les estimations. En effet, le résidu issu de la prévision ou de l'estimation prend en compte l'effet d'éventuelles variables exogènes. En réalité, une forte erreur peut provenir du fait que les gens ont tenu compte des conseils de Bison Futé et changé de comportement, ou alors d'une météorologie non favorable…. Tous ces événements sont impossibles à prévoir un an à l'avance et donc ne peuvent pas intervenir dans les critères qui permettent de définir les jours de l'historique semblables au jour j de prévision.

Représentation schématique de la construction de l'ensemble :



(b) Utilisation de l'ensemble

La première application sur E sera le calcul de la moyenne, de l'écart type et du coefficient de variation.

Ensuite le tracé de l'histogramme pourra montrer comment se répartissent les éléments de l'ensemble et la situation de la prévision dans l'ensemble. Une faible dispersion correspondra à un indice de confiance élevée dans la prévision. Par contre une forte dispersion nous amènera à conclure en une mauvaise prévision à priori.

- La moyenne

Par rapport à l'ensemble de prévisions E obtenu avec le système du pauvre, on peut calculer un intervalle de confiance de la moyenne de l'ensemble pour voir si la prévision se situe dans cet intervalle (ce qui peut nous donner une autre indication concernant la qualité de cette prévision).

Or, la loi de l'ensemble E n'est généralement pas une loi connue (et souvent diffère pour chaque prévision). De plus l'ensemble E contient N+1 éléments avec généralement N<50. Cette constatation nous amène à penser que du fait du faible nombre d'éléments de E, la moyenne n'est pas une statistique suffisamment robuste et ne représente pas réellement l'espérance de la loi. Une méthode aurait été d'estimer la densité de l'échantillon par des méthodes non paramétriques. Or, dans cette application nous ne sommes pas vraiment intéressés par l'obtention de la densité mais plus par l'obtention d'une moyenne fiable et d'un intervalle de confiance de cette moyenne. Nous avons donc décidé d'utiliser la méthode du bootstrap.

Rappel sur le bootstrap :

C'est une procédure de rééchantillonnage dont l'objectif est d'étudier les propriétés d'une statistique $T(X_1, X_2,\ldots, X_n, P)$ fondée sur un échantillon $(X_1, X_2,\ldots, X_n)$ d'une variable aléatoire X de loi P.

Soit une variable aléatoire X de loi P, de fonction de répartition (f.r.) F dont on possède un échantillon indépendant $E = (X_1, X_2,\ldots, X_n)$. L'idée est de ré-échantillonner de façon indépendante dans *E* et d'étudier le comportement d'une statistique $T(X_1, X_2,\ldots, X_n, F)$.

L'algorithme du bootstrap peut être résumé ainsi :

Phase 1 : *E* sert de population de base et est munie de la loi de probabilité empirique

$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ de f.r. F_n.

Phase 2 : Conditionnellement à P_n, on procède dans *E* à N tirages équiprobables avec remise ; $E^* = (X_1^*, X_2^*, \ldots, X_n^*)$ est l'échantillon ainsi obtenu , avec pour tout i=1 à N : il existe j, 1>j>n tel que $X_i^* = X_j$.

En pratique on prendra N=n.

Phase 3 : On approche le comportement de $T(E, F)$ par celui de $T(E^*, F_n) = T^*$; T^* est la statistique bootstrapée.

La phase 2 est répétée B fois (B relativement grand) engendrant B échantillons E_k^*, k=1…B, avec $E_k^*=(X_{1k}^*, X_{2k}^*, …., X_{Nk}^*)$. On observe donc B valeurs $T_k^*=T(E_k^*,F_n)$ de T.

Dans notre cas la statistique T correspond à la moyenne. Pour obtenir un intervalle de confiance de la moyenne, nous allons utiliser la méthode des percentiles de la distribution bootstrap d'une statistique. L'idée est d'utiliser les percentiles de l'histogramme du bootstrap pour définir les limites de l'intervalle de confiance.

On suppose qu'un échantillon E^* est généré et on note $\hat{\theta}^* = T^*$. En générant B échantillons, sur chaque échantillon on calcule $\hat{\theta}^*$. Soit $\hat{G}$ la fonction de distribution des $\hat{\theta}^*$. L'intervalle percentile 1-2α est défini par les α et 1−α percentiles de $\hat{G}$:

$$\left[\hat{\theta}_{\%,lo} - \hat{\theta}_{\%,up}\right] = \left[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)\right]$$

Par définition, $\hat{G}^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$ est le percentile de la distribution bootstrap.

Remarque : L'obtention d'un intervalle de confiance de la moyenne par le bootstrap est une information supplémentaire quant à la qualité de la prévision. Néanmoins grâce au modèle GLM, on peut obtenir un intervalle de confiance pour la prévision elle-même. (voir Annexe)
Cet intervalle est construit à partir des variables explicatives du modèle GLM et donc ne fait à aucun moment intervenir les débits réels de l'historique.
Les deux intervalles de confiance obtenus ne sont pas comparables. Le premier (par le bootstrap) est calculé par rapport aux données réelles du débit relatif, tandis que le deuxième fait référence aux variables calendaires (qualitatives).

- Recherche d'un débit relatif critique

Une autre possibilité est de donner un débit relatif critique. C'est-à-dire de chercher la valeur u_α pour laquelle $P(q_r(j) > u_\alpha) = 5\%$. Ce quantile correspond à une valeur du débit relatif critique qu'il ne faut pas dépasser (si par exemple, la prévision du jour j se trouve être supérieure à ce quantile, on pourra affirmer que cette prévision est vraiment mauvaise par rapport à l'ensemble).

Rappels sur les quantiles :

Les statistiques d'ordre servent tout particulièrement à construire d'autres statistiques dont le rôle en estimation et en théorie des tests est important.

Définition des p-quantiles d'une loi de probabilité sur R :
Soit X une variable aléatoire réelle de loi F. Soit $p \in]0;1[$. On appelle p-quantile de F tout réel $\zeta_p(F)$ vérifiant :

$$F(\zeta_p(F)-0) < p < F(\zeta_p(F))$$

Si la fonction F est continue strictement croissante, le p-quantile $\zeta_p(F)$ est unique et vérifie $F^{-1}(p)= \zeta_p(F)$ pour tout p compris entre zéro et un.
Sil existe un intervalle $I=[a;b[$ contenu dans R tel que
$\forall x \in [a;b[\ F(x)=p\ \ p \in]0;1[$ alors tout point de l'intervalle I est un p-quantile. C'est le seul cas où il n'y a pas unicité du quantile.
Si F est discontinue en x_0 c'est-à-dire $F(x_0-0)<F(x_0)$ alors $\zeta_p(F)$ est unique et égal à x_0 pour tout $p \in]F(x_0-0);F(x_0)[$.

Définition du p-quantile empirique :
Soit X_1, X_2,…. X_n un échantillon d'une loi F. Soit $p \in]0;1[$. On appelle p-quantile empirique et on note $\zeta_p(F_n)$, un p-quantile de la fonction de répartition empirique.

Si F est continue et si p n'appartient pas à l'ensemble {1/n, …., (n-1)/n} alors le p-quantile empirique est l'une des composantes du vecteur des statistiques d'ordre $X_{(.)}$. A l'opposé un k/n-quantile empirique est une combinaison linéaire convexe de $X_{(k)}$ et $X_{(k+1)}$. Dans ce dernier cas, pour éliminer tout ambiguïté, nous prenons pour k/n-quantile le milieu de l'intervalle $[X_{(k)} ; X_{(k+1)}]$.
Nous avons donc

$$\zeta_p(F_n)= X_{(k+1)} \text{ si } p \in]k/n;(k+1)/n[$$
$$(X_{(k)} +X_{(k+1)})/2 \text{ si } p=k/n$$

Les p-quantiles empiriques sont des estimateurs convergents des p-quantiles de F. Ils peuvent aussi servir de « bons estimateurs » d'un paramètre de localisation. De même les différences $\zeta_q(F_n)- \zeta_p(F_n)$ avec q < p servent d'estimateurs de paramètres d'échelle.

Pour rendre plus robuste la valeur du quantile, on peut comme pour la moyenne utiliser des échantillons bootstrapés et regarder la distribution des quantiles calculés sur chaque échantillon ce qui nous permet d'obtenir un intervalle de confiance pour la moyenne des quantiles bootstrapée (statistique qui converge plus vite vers la vraie valeur du quantile de l'ensemble final).

Une autre manière d'obtenir également ce débit critique est de faire une estimation non paramétrique de la densité de l'ensemble. Cette méthode un peu plus compliquée à mettre en œuvre a l'avantage de donner des résultats non empiriques.

2) Application à la prévision des débits relatifs de l'année 98 :

Nous présenterons dans ce paragraphe les résultats obtenus par la prévision d'ensemble appliqués au capteur de Saint Arnoult dans le sens Paris Province. Pour chaque ensemble, on essaiera d'associer à la prévision du modèle un indice de confiance permettant ainsi de la qualifier à priori. Celui-ci est déduit de l'analyse de l'ensemble de prévision. Il faudra regarder à la fois si cet ensemble est homogène d'un point de vue calendaire, la dispersion de l'histogramme (notamment s'il existe des classes équiprobables), ainsi que les statistiques associées à l'ensemble.
Cet indice prendra, de la même façon que pour la météorologie la forme d'un chiffre entre un et quatre avec les caractéristiques suivantes:

1 confiance nulle : prévision mal placée et/ou histogramme plat

2 faible confiance : prévision bien placée mais au moins deux classes équiprobables et distancées en valeur, significatives pour notre application (une différence de 0.1 sur le taux relatif représente environ une différence de 3 000 véhicules sur le débit)

3 confiance normale : prévision bien placée et histogramme asymétrique mais non plat

4 bonne confiance : prévision bien placée et histogramme proche d'une normale

Un tableau récapitulatif permettra ensuite de comparer cet indice (qualificatif à priori) à l'erreur calculée à posteriori pour certains jours (ordinaires ou exceptionnels).

(a) Prévision d'un jour ordinaire :

On veut construire l'ensemble de prévisions pour le dimanche 27/09/98 qui est un jour ordinaire. Pour cela on applique le système du pauvre avec pour historique tous les dimanche ordinaires du passé c'est-à-dire des années 1987 à 1997.

En appliquant la méthode du système du pauvre on obtient un ensemble final assez homogène du point de vue calendaire car comprenant des dimanches des mois de septembre, octobre, décembre, mars, avril…. La moyenne de l'ensemble vaut 0.6857 avec un écart type de 0.0472.

Le tableau ci-après donne les jours de l'ensemble final avec en première ligne le jour de prévision et donc la prévision de contrôle.

ensemble prévision débit relatif du 27/09/98 (jour ordinaire)

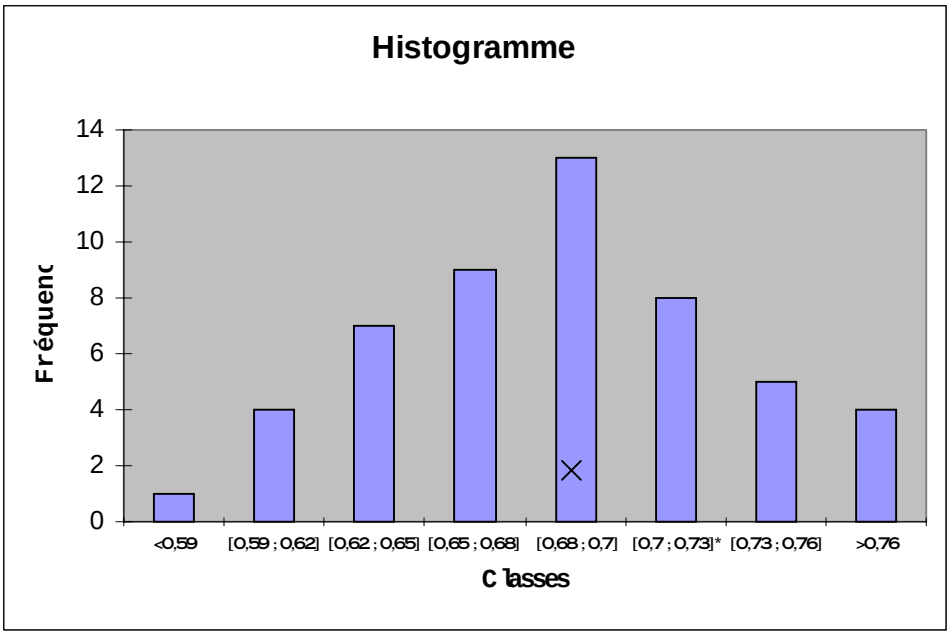
jour	mois	année	Type de jours	débit relatif
27	9	98	Dimanche	0,683
1	12	91	Dimanche	0,607
17	5	87	Dimanche	0,665
16	*5*	*93*	Dimanche	*0,781*
14	5	95	Dimanche	0,664
12	5	96	Dimanche	0,696
13	5	90	Dimanche	0,601
5	*5*	*96*	Dimanche	*0,681*
11	*12*	*94*	Dimanche	*0,670*
13	*12*	*87*	Dimanche	*0,765*
10	*12*	*95*	Dimanche	*0,732*
12	12	93	Dimanche	0,630
11	12	88	Dimanche	0,641
7	12	97	Dimanche	0,647
8	*12*	*96*	Dimanche	*0,662*
9	*12*	*90*	Dimanche	*0,667*
10	*12*	*89*	Dimanche	*0,684*
8	12	91	Dimanche	0,645
7	4	91	Dimanche	0,594
8	*9*	*96*	Dimanche	*0,746*
7	*9*	*97*	Dimanche	*0,787*
10	9	95	Dimanche	0,659
11	*9*	*94*	Dimanche	*0,744*
12	9	93	Dimanche	0,702
10	*9*	*89*	Dimanche	*0,710*
13	9	87	Dimanche	0,683

11	*9*	*88*	Dimanche	*0,708*
14	*10*	*90*	Dimanche	*0,767*
12	10	97	Dimanche	0,667
15	*10*	*95*	Dimanche	*0,707*
13	*10*	*96*	Dimanche	*0,722*
15	*10*	*89*	Dimanche	*0,724*
13	10	91	Dimanche	0,698
31	*3*	*96*	Dimanche	*0,698*
23	4	89	Dimanche	0,631
24	3	91	Dimanche	0,615
29	3	87	Dimanche	0,626
27	*3*	*94*	Dimanche	*0,689*
24	*3*	*96*	Dimanche	*0,673*
28	*3*	*93*	Dimanche	*0,691*
25	3	90	Dimanche	0,603
26	3	95	Dimanche	0,637
23	3	97	Dimanche	0,658
9	*2*	*97*	Dimanche	*0,735*
15	*9*	*91*	Dimanche	*0,721*
15	*9*	*96*	Dimanche	*0,708*
14	*9*	*97*	Dimanche	*0,713*
28	*9*	*97*	Dimanche	*0,753*
29	9	91	Dimanche	0,681
30	*9*	*90*	Dimanche	*0,699*
29	*9*	*96*	Dimanche	*0,704*

moyenne	0,686
écart type	0,047
coef variation	0,069

L'ensemble est assez homogène d'un point de vue calendaire car même si plusieurs mois apparaissent, les mois d'été (juin, juillet et août) sont absents. La saison n'intervient pas dans le cadre de la prévision pour un jour ordinaire puisque les déplacements sont essentiellement des trajets domicile travail. Néanmoins comme le jour de prévision est un dimanche d'été, on peut penser que les personnes se déplacent davantage le week-end que durant l'année.

L'histogramme ci après montre une faible dispersion.

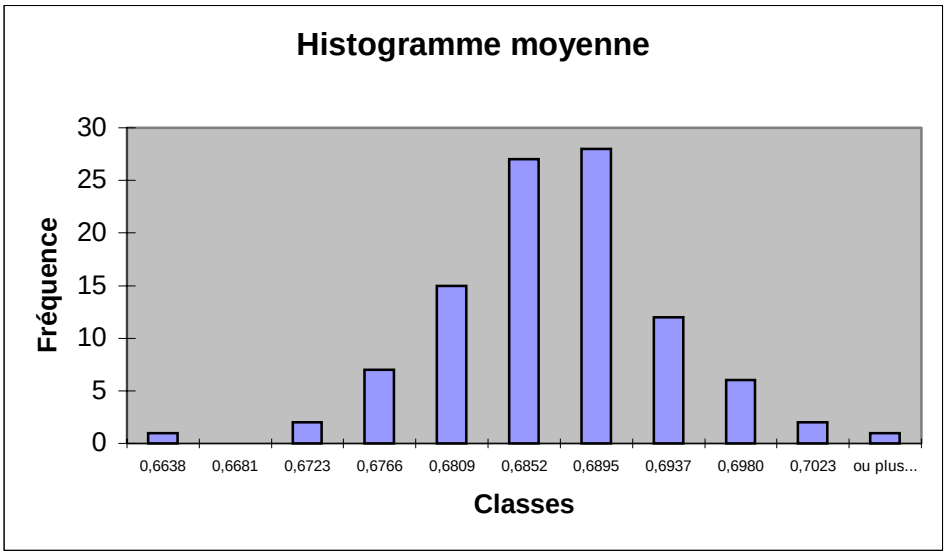


D'après la forme classique de l'histogramme, on remarque que l'ensemble de prévisions a une dispersion assez homogène autour de la valeur moyenne et donc aussi autour de la prévision puisque les deux valeurs sont similaires.

A priori on a, d'après l'histogramme, une bonne confiance en la prévision donnée par le modèle GLM.

Avec le bootstrap en générant 100 échantillons à partir de l'échantillon initial (hormis la prévision) on obtient la moyenne bootstrapée égale à 0.6849.L'intervalle de confiance à 95% basée sur les percentiles de cette valeur est [0.6731 ; 0.6954].

L'histogramme des moyennes étant le suivant



La prévision appartient à l'intervalle de confiance de la moyenne bootstrapée.

De plus, en considérant l'écart type donné avec la prévision par la méthode GLM, on obtient un intervalle à 95% pour la prévision qui est [0.6506 ; 0.7162].

En conclusion on peut dire que d'après l'ensemble de prévisions (basé sur les débits relatifs réels du passé) la confiance en la prévision du modèle GLM est bonne. On pourrait lui associer un indice de confiance égal à 4 c'est-à-dire une très bonne confiance.

(b) Un autre jour ordinaire :

Cette fois, la prévision porte sur un vendredi, le 06/11/1998. C'est un jour qualifié d'ordinaire si l'on regarde d'un point de vue variables calendaires explicatives du modèle GLM. Néanmoins il convient de préciser que ce jour n'est pas tout à fait ordinaire dans le sens où, en plus des trajets domicile travail, se rajoutent les départs en week-end d'autant plus importants que le mercredi suivant est le 11 novembre donc férié. On garde la notion de jour ordinaire pour construire l'ensemble final mais on traitera ce jour avec méfiance.

La construction de l'ensemble de prévisions via le système du pauvre donne l'ensemble ci dessous :

ensemble prévision débit relatif du 06/11/98 (jour ordinaire)

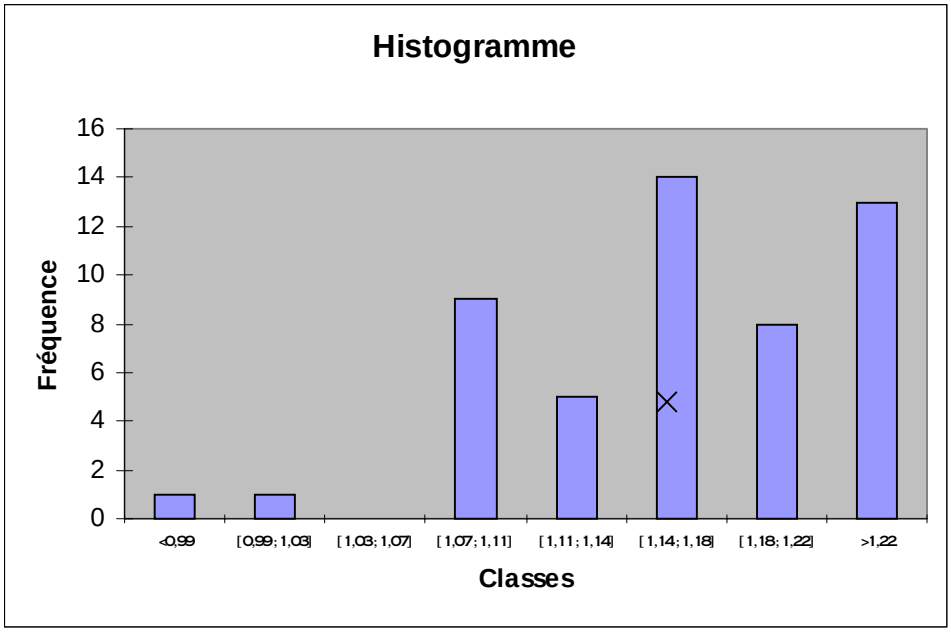
jour	mois	année	Type de jours	ensemble
6	11	98	Vendredi	1,179
1	12	95	Vendredi	1,223
1	12	89	Vendredi	1,187
4	12	87	Vendredi	1,135
10	5	96	Vendredi	1,211
26	11	93	Vendredi	1,106
22	11	91	Vendredi	1,116
25	11	94	Vendredi	1,153
22	11	96	Vendredi	1,010
25	11	88	Vendredi	1,151
21	11	97	Vendredi	1,178
24	11	89	Vendredi	1,240
23	11	90	Vendredi	1,164
27	11	87	Vendredi	1,178
24	11	95	Vendredi	1,259
11	3	88	Vendredi	1,108
12	3	93	Vendredi	1,241
9	3	90	Vendredi	1,203
13	3	87	Vendredi	1,122
7	3	97	Vendredi	1,186
10	3	89	Vendredi	1,149
21	3	97	Vendredi	1,204
27	3	87	Vendredi	0,997
22	3	91	Vendredi	1,178
23	3	90	Vendredi	1,257
22	3	96	Vendredi	1,190
24	3	95	Vendredi	1,242
25	3	94	Vendredi	1,183
26	3	93	Vendredi	1,226
18	3	94	Vendredi	1,238
18	3	88	Vendredi	1,085
15	3	91	Vendredi	1,164
19	3	93	Vendredi	1,248

17	*3*	*95*	*Vendredi*	*1,162*
16	*3*	*90*	*Vendredi*	*1,242*
14	*3*	*97*	*Vendredi*	*1,170*
17	3	89	*Vendredi*	1,080
20	3	87	*Vendredi*	1,119
21	4	89	*Vendredi*	1,174
22	*4*	*88*	*Vendredi*	*1,241*
19	11	93	*Vendredi*	1,092
18	11	88	*Vendredi*	1,122
18	11	94	*Vendredi*	1,091
17	*11*	*89*	*Vendredi*	*1,235*
17	*11*	*95*	*Vendredi*	*1,171*
15	11	96	*Vendredi*	1,075
16	*11*	*90*	*Vendredi*	*1,233*
15	11	91	*Vendredi*	1,086
20	*11*	*87*	*Vendredi*	*1,203*
14	11	97	*Vendredi*	1,076
3	*5*	*96*	*Vendredi*	*1,200*

moyenne	1,166
écart type	0,063
coef variation	0,054

D'un point de vue calendaire, beaucoup de vendredis du mois de mars sont présents ce qui s'explique par le fait que les gens commencent à partir plus souvent en week-end à cette époque.

L'histogramme de l'ensemble :



La croix représente la prévision.

A priori, la prévision obtenue par le modèle n'est pas très réaliste. En effet, d'après l'histogramme construit avec des débits relatifs réels de l'historique il y a plus de 50% d'éléments qui se trouvent être supérieurs à la valeur réelle. Cette constatation montre qu'en considérant l'historique du jour à prédire, le débit relatif s'est trouvé souvent supérieur à la valeur prévue par le modèle GLM.

De plus en regardant de plus près les fréquences calculées sur l'ensemble final on remarque qu'en ce qui concerne les probabilités empiriques on a la propriété suivante :

$$P(q_r(j) \in [1.14;1.18]) = \frac{14}{50} \approx P(q_r(j) > 1.22) = \frac{13}{50}$$

Il y a de fortes chances que le débit réel puisse être nettement supérieur à la valeur prédite.

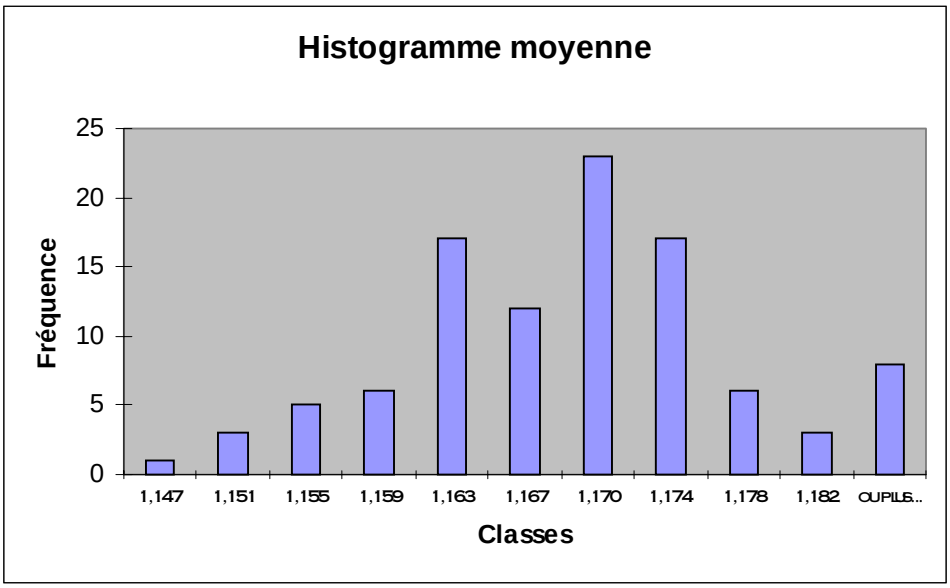
Pour l'intervalle de confiance de la prévision, l'écart type calculé sur l'historique de la prévision est de 0,02232. Donc l'intervalle de confiance est [1.1352 ; 1.2226].

Si on considère la moyenne, elle prend une valeur très proche de la valeur prédite, compte tenu de la dispersion de l'ensemble de prévision.

Pour rendre la moyenne plus robuste nous avons généré 100 échantillons bootstrapés à partir de l'ensemble de prévisions (auquel on a retiré la prévision du modèle).

Pour chaque échantillon bootstrapé on a calculé la moyenne, puis on a ainsi obtenu un vecteur de moyennes (de dimension 101 en incluant la moyenne de l'ensemble de prévisions).

On trace l'histogramme des moyennes.



On peut donc calculer sur cet échantillon la moyenne qui étant données les propriétés du bootstrap converge plus vite vers la vraie valeur de l'espérance de l'ensemble final. De plus, grâce à la méthode des percentiles on construit un intervalle de confiance pour cette moyenne.

moyenne bootraspée
1,167

intervalle de confiance moyenne bootstrapée
[1,1496 ; 1,1834]

La prévision issue du modèle appartient à cet intervalle ainsi que la moyenne calculée sur l'ensemble ce qui confirme que ces deux valeurs sont très proches.

En utilisant la même méthode que pour la moyenne, on calcule un seuil critique empirique, d'abord en ne considérant que l'ensemble final puis en le rendant plus robuste par la méthode du bootstrap.

quantile empirique 5% calculé sur l'ensemble final
1,245

moyenne quantile empirique bootstrapé
1,247

intervalle de confiance de la moyenne quantile empirique bootstrapé
[1,2414 ; 1,2588]

L'analyse ci-dessus montre que la prévision faite par le modèle est mauvaise et que par conséquent on a une confiance très mitigée en cette valeur. Toutefois la dispersion de l'ensemble final a permis quand même d'anticiper la forte probabilité que le débit relatif réel soit nettement supérieur à celui que l'on envisage. L'alternative $\{q_r(j) > 1.22\}$ a une forte probabilité de se réaliser. Au vu de ces différentes constatations, on peut donc associer un indice de confiance égal à 2 (faible confiance) à la prévision.

En effet, même si la moyenne de l'ensemble et la prévision ont des valeurs très similaires il existe néanmoins une forte probabilité que l'événement alternatif se réalise. On associe donc une faible confiance à cette prévision car la dispersion de l'ensemble permet de voir dans quel sens un phénomène alternatif peut se produire.

Ici donner un indice de confiance égal à 2 (faible confiance) permet de détecter ce qui semble être une mauvaise prévision. Mais surtout ce qui importe est que cette méthodologie permet de dire à priori si la prévision est une sous estimation ou une sur estimation du débit relatif réel.

On vient de voir deux exemples de jours ordinaires. Au premier l'indice de confiance associé à la prévision était de 4, au suivant de 2. Il est important de signaler que dans la plupart des cas, les indices de confiance que l'on peut associer à ces jours dits ordinaires sont de 4 . Cette remarque va dans le même sens que les constatations faites lors de l'analyse des résultats du modèle GLM, qui concluaient à une bonne prévision pour les jours ordinaires.

De plus, il faut aussi savoir que les débits relatifs des jours ordinaires sont nettement inférieurs que ceux des jours exceptionnels. Une erreur relative sur ces taux n'entraîne pas forcément une erreur conséquente en terme de débit journalier (c'est-à-dire de nombre de véhicules). En ce qui concerne le débit, une erreur de prévisions devient trop grande lorsque la différence entre le débit réel et le débit prévu dépasse les 5 000 véhicules.

(c) Prévision pour un jour exceptionnel de l'été :

On cherche l'ensemble pour un vendredi , le 07/08/98. Ce vendredi est l'un des premiers du mois d'août donc on s'attend à beaucoup de départs en vacances.

Pour la construction, on a appliqué le système du pauvre en ne retenant dans l'historique que les vendredi exceptionnels des mois de juin à septembre pour avoir une homogénéité.

Pour ce week-end, on obtient comme ensemble final :

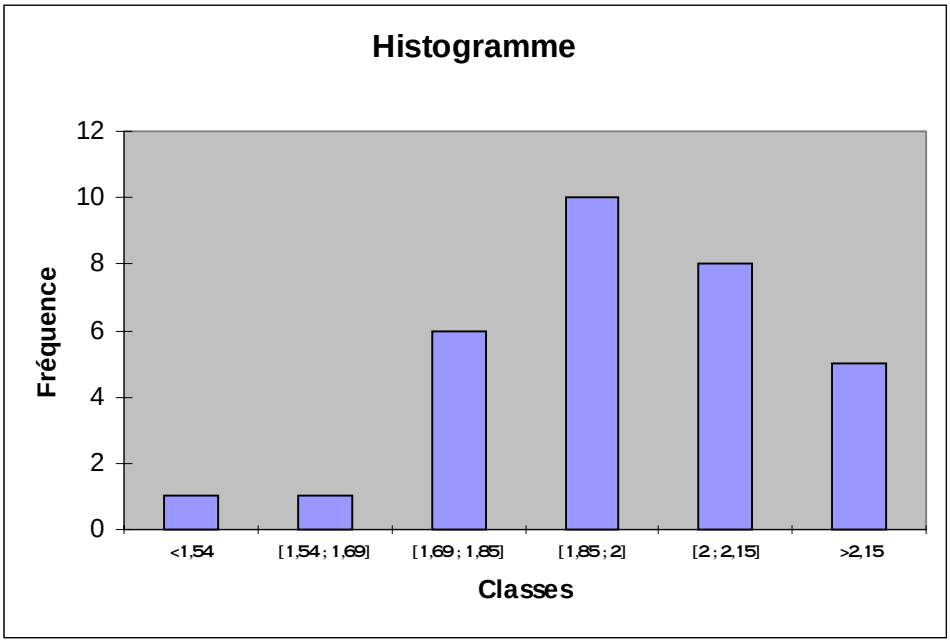
ensemble prévision débit relatif 07/08/98 vendredi code calendaire 1

jour	mois	année	Type de jour	code calendaire	Débit relatif
7	8	98	Vendredi	0	1,873
12	8	94	Vendredi	5	2,145
7	9	90	Vendredi	0	1,541
16	7	93	Vendredi	0	2,093
15	*8*	*97*	Vendredi	*6*	*1,651*
14	*8*	*87*	Vendredi	*5*	*2,106*
17	7	87	Vendredi	0	1,988
30	*6*	*89*	Vendredi	*1*	*2,307*
27	6	97	Vendredi	1	1,871
29	*6*	*90*	Vendredi	*1*	*2,158*
30	*6*	*95*	Vendredi	*2*	*2,057*
21	*7*	*89*	Vendredi	*0*	*2,062*
19	*7*	*96*	Vendredi	*0*	*2,161*
18	7	97	Vendredi	0	1,993
20	7	90	Vendredi	0	1,780
19	*7*	*91*	Vendredi	*0*	*2,207*
21	7	95	Vendredi	0	1,883
22	7	88	Vendredi	0	1,976
22	*7*	*94*	Vendredi	*0*	*2,022*
23	*7*	*93*	Vendredi	*0*	*2,132*
5	*7*	*91*	Vendredi	*1*	*2,262*
15	7	88	Vendredi	0	1,728
15	*7*	*94*	Vendredi	*0*	*1,819*
1	*7*	*94*	Vendredi	*0*	*2,067*
2	*7*	*93*	Vendredi	*0*	*1,982*
13	8	93	Vendredi	5	1,846
8	8	97	Vendredi	0	1,887
10	*8*	*90*	Vendredi	*0*	*1,968*
9	8	91	Vendredi	0	1,863
9	8	96	Vendredi	0	1,805
8	7	88	Vendredi	0	1,803

moyenne	1,969
écart type	0,179
coef variation	0,091

Déjà, on constate que la moyenne de l'ensemble diffère sensiblement de la valeur prédite avec une différence d'environ 0.1 ce qui correspond à environ 3 500 véhicules, valeur qui est significative.

La répartition de l'ensemble final donne l'histogramme suivant :



La dispersion autour de la prévision laisse penser qu'il y a de fortes chances que l'on soit dans un cas de sous estimation. En effet, en regardant de plus près les fréquences associées à cet histogramme on a :

$$P(q_r > 2) = \frac{13}{30} > P(q_r \in [1.85;2]) = \frac{10}{30}$$

En utilisant la méthode du bootstrap, pour obtenir à la fois une moyenne plus robuste et une valeur critique empirique, on arrive aux statistiques suivantes :

moyenne bootraspée
1,970

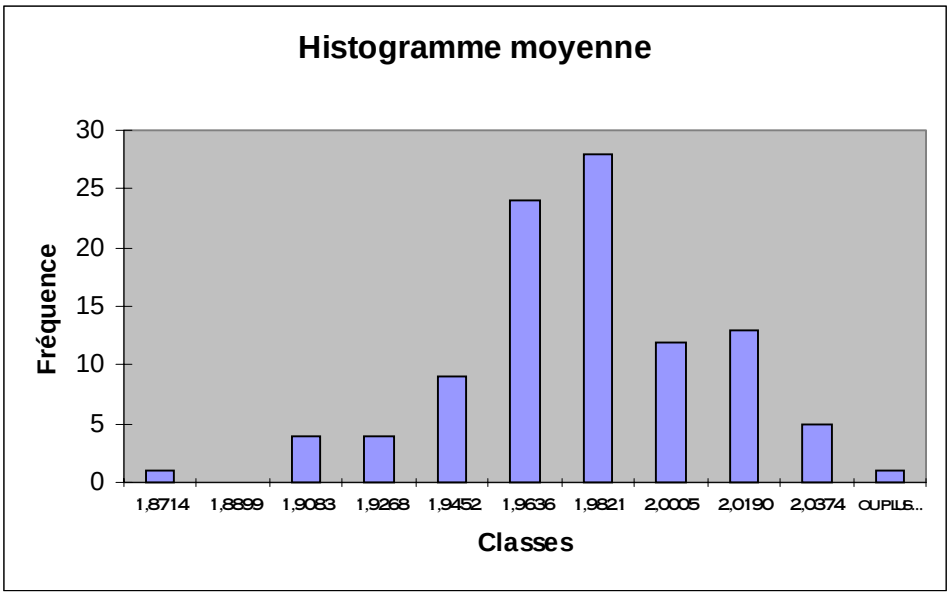
intervalle de confiance moyenne bootstrapée
[1,9026 ; 2,0281]

quantile empirique 5%
2,235

moyenne quantile empirique bootstrapé
2,230

intervalle de confiance de la moyenne quantile empirique boot
[2,1401 ; 2,3073]

Le tracé de l'histogramme des moyennes calculées sur les 100 échantillons bootstrapés confirme la remarque précédente, c'est-à-dire qu'il y a de fortes chances pour que le débit réel soit plus important à ce que l'on prévoit.



L'intervalle de confiance de la moyenne bootstrapée prend comme valeur supérieure 2.0281. De plus la répartition de l'histogramme ci dessus montre que beaucoup de moyenne se trouvent être supérieures à 2.

L'indice de confiance associé à cette prévision est donc de 2, à savoir une faible confiance. Néanmoins, la répartition de l'ensemble final n'est pas complètement aléatoire puisqu'elle montre clairement que la prévision déduite du modèle sous estime la valeur réelle du débit relatif. En considérant la moyenne de l'ensemble final comme nouvelle prévision, on peut s'attendre à corriger du moins en partie cette sous estimation. Il semble donc plus logique de s'attendre à un débit relatif autour de 1.97 au lieu de 1.87.

(d) Prévision pour un autre jour exceptionnel de l'été :

On cherche à construire l'ensemble pour le 13/07/98 qui est un lundi (lundi de départ avant le pont du 14 juillet). Ce jour du point de vue des variables explicatives du modèle GLM est considéré comme exceptionnel (veille de pont) or ce n'est pas tout à fait le cas. En effet, les gens sont partis essentiellement le vendredi. Néanmoins ce n'est tout de même pas un jour ordinaire puisque certaines personnes peuvent quitter Paris en sachant que le lendemain est férié. Le débit risque de se rapprocher de celui d'un jour ordinaire chargé.

ensemble prévisions débit relatif 13/07/98 (lundi code calendaire 10)

jour	mois	année	Type de jours	code calendaire	débit relatif
13	7	98	Lundi	0	0,967
18	*8*	*97*	*Lundi*	*0*	*0,945*
22	*7*	*96*	*Lundi*	*0*	*1,089*
22	*7*	*91*	*Lundi*	*0*	*1,085*
21	7	97	*Lundi*	0	1,059
23	7	90	*Lundi*	0	1,048

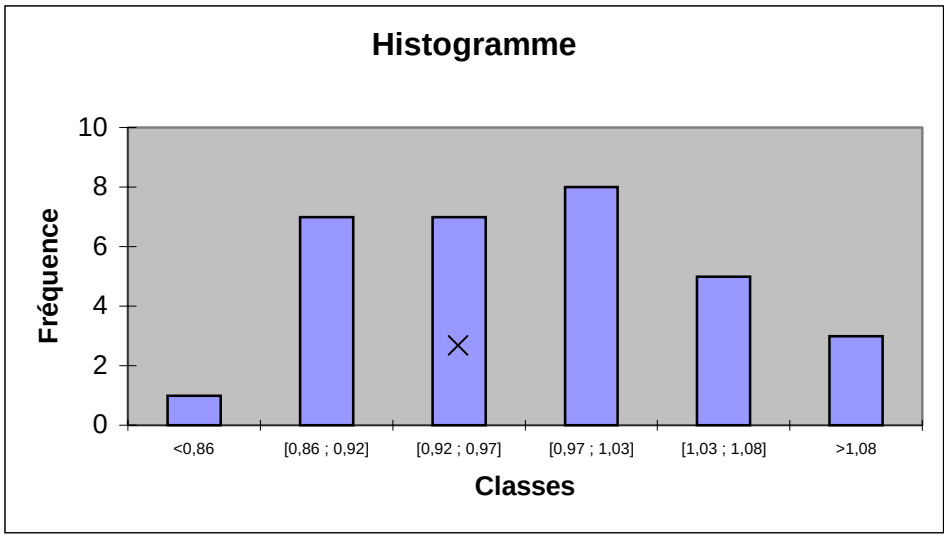
20	7	87	*Lundi*	0	1,005
19	*7*	*93*	*Lundi*	*0*	*1,139*
14	*7*	*97*	*Lundi*	*10*	*0,886*
19	8	96	*Lundi*	0	0,898
19	*8*	*91*	*Lundi*	*0*	*0,902*
6	*9*	*93*	*Lundi*	*7*	*0,961*
5	*9*	*94*	*Lundi*	*7*	*1,021*
4	*9*	*89*	*Lundi*	*7*	*0,959*
4	9	95	*Lundi*	7	0,909
2	9	96	*Lundi*	7	0,864
5	*9*	*88*	*Lundi*	*7*	*0,980*
7	*9*	*87*	*Lundi*	*7*	*0,929*
13	7	87	*Lundi*	0	1,000
13	8	90	*Lundi*	0	0,941
12	8	91	*Lundi*	0	0,980
11	*8*	*97*	*Lundi*	*0*	*1,009*
12	8	96	*Lundi*	0	0,997
8	*8*	*88*	*Lundi*	*0*	*0,978*
8	*8*	*94*	*Lundi*	*0*	*1,082*
9	*8*	*93*	*Lundi*	*0*	*1,079*
10	8	87	*Lundi*	0	0,896
9	9	91	*Lundi*	7	0,943
1	9	97	*Lundi*	0	0,891
16	*8*	*93*	*Lundi*	*0*	*1,075*
17	8	87	*Lundi*	0	0,866

moyenne	0,980
écart type	0,075
coef variation	0,076

Le code calendaire correspond à un qualificatif pour les jours exceptionnels par exemple, rentrée des classes, premier jour de pont…

Au niveau purement calendaire, il n'y a pas d'aberration puisque les jours obtenus dans l'ensemble se ressemblent. L'ensemble apparaît donc homogène. La moyenne est assez proche de la prévision.

L'histogramme de l'ensemble est présenté ci dessous :



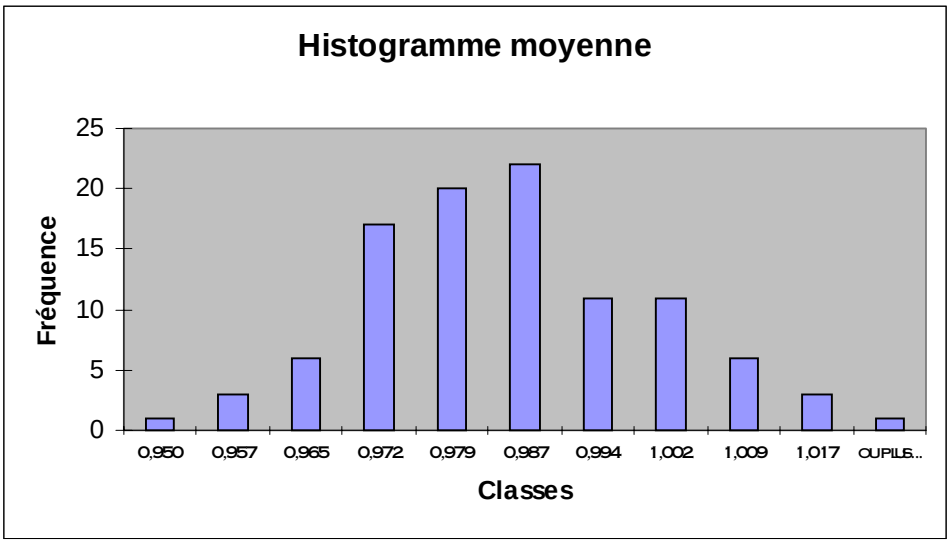
La dispersion autour de la moyenne (et donc autour de la prévision) est assez forte. Il n'y a pas vraiment pas d'homogénéité dans la forme de l'histogramme.

Il y a trois intervalles où la probabilité empirique que le débit relatif appartienne à ces intervalles est quasiment la même. On a une sorte d'histogramme plat avec donc une équiprobabilité. Il semble difficile alors de situer la valeur la plus probable du débit relatif.

L'intervalle de confiance de la prévision obtenu par le modèle GLM est [0,8568 ; 1.0781].

On a généré 100 échantillons bootstrapés afin d'obtenir un intervalle de confiance pour la moyenne de l 'ensemble.

L'histogramme des moyennes des 100 échantillons est



On obtient différentes valeurs associées à la méthode :

moyenne bootraspée
0,982

intervalle de confiance moyenne bootstrapée
[0,9525 ; 1,0092]

En faisant la même étude sur le quantile empirique à 5% on a les résultats suivants :

quantile empirique 5%
1,0870

moyenne quantile empirique bootstrapé
1,093

intervalle de confiance de la moyenne quantile empirique boot
[1,0774 ; 1,1388]

La forme de l'histogramme de l'ensemble de prévisions est vraiment dispersée et surtout équiprobable. Même si la moyenne de l'ensemble et la prévision sont proches, on ne peut se fier à cette prévision puisque la répartition autour de cette moyenne est trop large (il y a trop de possibilités qui diffèrent de la prévision).

L'indice de confiance que l'on peut attribuer à cette prévision est donc de 1 c'est-à-dire que la confiance que l'on peut associer à cette prévision est nulle. Pour ce type de jours où la

confiance en la prévision est nulle, on peut donner grâce à l'ensemble un seuil inférieur permettant ainsi de trouver une valeur empirique u_α telle que :

$P(q_r(j) > u_\alpha) = 95\%$.

Pour cela on utilise le quantile empirique calculé sur l'ensemble final mais pour rendre plus robuste sa valeur on applique la méthode du bootstrap. Les résultats obtenus sont les suivants :

quantile empirique 95% calculé sur l'ensemble final
0,876

moyenne quantile empirique bootstrapé
0,878

intervalle de confiance
[0,8640 ; 0,8984]

En définitive, même si on associe à la prévision une confiance nulle, il est possible de donner un seuil inférieur critique et de conclure que
$P(q_r > 0.877) = 95\%$.
On n'est pas en mesure de savoir la valeur exacte du débit relatif réel (ni celle qui s'en rapproche le plus) par contre on peut affirmer avec une forte probabilité que le taux journalier réel sera supérieur à 0.877.

(e) Tableau récapitulatif et comparatif pour un certain nombre de jours :

date	type de jours	ordinaire/ exceptionnel	débit relatif prévu	indice de confiance	débit relatif réel	erreur relative	remarques
27/02/98	vendredi	ordinaire	0,9973	2	1,2058	17,29	on détecte la sous estimation à priori
05/03/98	jeudi	ordinaire	0,7596	3	0,778	2,37	
12/03/98	jeudi	ordinaire	0,7657	4	0,766	0,04	
18/03/98	mercredi	ordinaire	0,7264	4	0,7178	-1,20	
26/06/98	vendredi	exceptionnel	2,0344	1	1,6399	-24,06	forte dispersion (classé comme jour exceptionnel mais ressemble plus à un jour ordinaire)
03/07/98	vendredi	exceptionnel	2,211	2	2,109	-4,84	forte dispersion
10/07/98	vendredi	exceptionnel	2,7754	2	2,7347	-1,49	non homogénéité calendaire dans l'ensemble final
13/07/98	lundi	exceptionnel	0,9674	1	1,2189	20,63	équiprobabilité
27/07/98	lundi	exceptionnel	1,2683	2	1,0809	-17,34	on détecte la sur estimation à priori
03/08/98	lundi	exceptionnel	1,3537	2	1,6385	17,38	on détecte la sous estimation à priori
07/08/98	vendredi	exceptionnel	1,8732	2	2,0961	10,63	on détecte la sous estimation à priori
09/08/98	dimanche	exceptionnel	1,0478	2	1,1912	12,04	on détecte la sous estimation à priori
16/08/98	dimanche	exceptionnel	0,7459	1	0,931	19,88	équiprobabilité

17/08/98	lundi	exceptionnel	0,9078	1	1,0243	11,37	équiprobabilité
29/08/98	samedi	exceptionnel	1,2258	1	1,2352	0,76	forte dispersion
27/09/98	dimanche	ordinaire	0,6834	4	0,6795	-0,57	
06/11/98	vendredi	ordinaire	1,1789	2	1,3695	13,92	on détecte la sous estimation à priori

Ce tableau présentant quelques comparaisons entre la vision à priori et celle à posteriori montre qu'il y a dans la majorité des cas concordants. Les jours où l'on détecte une mauvaise prévision à priori alors qu'elle s'avère juste sont des jours où l'ensemble final n'avait souvent pas une homogénéité d'un point de vue calendaire et où dans l'ensemble des jours semblables de l'historique beaucoup de scénario diverse ont eu lieu.

Néanmoins cette méthodologie de prévision d'ensemble permet souvent de détecter les cas de sous estimation (ou sur estimation). Il est rare qu'on arrive à un indice de confiance fort (égal à 4 ou 5) et que la prévision s'avère réellement mauvaise. Les fois où cela se produit correspondent à des jours où la variation de débit est imprévisible (les gens ont écouté les conseils de Bison Futé, ou un accident est apparu).

Troisième partie :
Simulation du trafic routier

On cherche dans cette partie à étudier les concepts de simulation du trafic pour voir si la méthode d'évaluation a priori des résultats utilisée en météorologie peut s'appliquer.

I. Présentation générale de la simulation du trafic routier

Les modèles permettent de donner une représentation simplifiée de la réalité sous forme de lois (c'est-à-dire de variables et de relations entre ces variables) et sont destinés aussi bien à améliorer la connaissance de cette réalité qu'à être partie intégrante d'un processus de contrôle. L'objectif des modèles est alors de tester des hypothèses d'évolution, d'évaluer des stratégies de commande ou l'influence d'un paramètre sur le comportement d'ensemble. La simulation est un processus de résolution du modèle, c'est-à-dire le calcul des états successifs. Un même modèle peut faire l'objet de divers modes de résolution, et un outil de simulation intègre éventuellement plusieurs modèles.
A partir d'un ensemble de conditions initiales (état du réseau au début de l'étude) et de conditions aux limites (demande en entrée du réseau, contraintes en sortie du réseau, incidents…) le modèle doit permettre de déterminer l'évolution des variables.

1) L'écoulement du trafic

Il existe essentiellement trois grandes catégories de modèles en trafic :
- Modèles microscopiques
- Modèles macroscopiques
- Modèles mésoscopiques

(a) Les modèles microscopiques

Ces modèles gèrent le trafic en individualisant chaque mobile. On s'intéresse aux variables individuelles caractérisant l'état de chaque véhicule : accélération, vitesse, position. Les lois de comportement et d'interaction (lois de poursuite) simulent la trajectoire de chacun de ces mobiles en fonction de leur environnement. Ce modèle de poursuite définit l'accélération (positive ou négative) de chaque véhicule. Lorsque sa distance au véhicule qui le précède est supérieure à une certaine valeur, le conducteur cherche à atteindre une vitesse maximum dite vitesse désirée. En dessous de cette valeur, l'accélération est fonction de différentes variables (la différence des vitesses, la distance inter véhiculaire, la vitesse du véhicule étudié…).

Expressions des lois de poursuite :
Il existe deux formes d'expression de lois de poursuite :

- La première est une forme additive. L'accélération du véhicule suiveur est décrite par une combinaison linéaire de la vitesse relative, la distance intervéhiculaire et sa vitesse absolue.

$$x''_n(t+T) = \alpha\,(x'_{n-1}(t) - x'_n(t)) + \beta\,(x_{n-1}(t) - x_n(t)) + \gamma\, x'_n(t)$$

où x_n et x_{n-1} désignent des positions, par rapport à un même point de référence, des véhicules n et n-1 et α, β, γ sont des coefficients de pondération liés à la sensibilité des conducteurs.

- La deuxième forme est non linéaire et l'accélération du véhicule suiveur dépend essentiellement de la vitesse relative. La sensibilité du conducteur est alors plus ou moins marquée selon sa vitesse individuelle et la distance qui le sépare du véhicule précédent.

$$x''_n(t+T) = \frac{c(x'_{n-1}(t) - x'_n(t)) * (x'_n(t))^m}{(x_{n-1}(t) - x_n(t))}$$

La description du modèle doit s'accompagner d'un modèle d'injection individuelle de véhicules fondé par exemple sur une loi de distribution statistique. Les caractéristiques individuelles des véhicules, vitesse désirée, peuvent elles-mêmes faire l'objet d'une distribution. Une description microscopique rend également nécessaire un modèle pour reproduire les dépassements, les changements de file.

L'existence d'un certain nombre d'aspects stochastiques rend donc non représentatifs les résultats obtenus à partir d'un seul calcul. Ils ne représentent en effet qu'une réalisation unique d'un phénomène aléatoire. La prise en compte de la distribution autour d'une valeur moyenne ne peut être faite que si un nombre significatif de valeurs définit un ensemble de conditions initiales (et non pas une seule).

(b) Les modèles macroscopiques

Ces modèles considèrent le trafic comme un écoulement continu. Les lois utilisées (dérivées de l'hydrodynamique) sont des équations liant les variables entre elles et des lois de propagation (ondes de choc). Le trafic est donc représenté comme un flux homogène, caractérisé par des grandeurs moyennes qui sont les variables d'état du modèle : concentration (C :répartition de véhicules dans l'espace en véhicules par mètre)), débit (Q : nombre de véhicules passés en un point lors d'un certain intervalle de temps), vitesse (V, vitesse moyenne du flot de véhicules). Le débit permet de définir la description de la demande. Trois principales équations sont utilisées.

La première équation est celle de conservation des masses, principe qui s'énonce ainsi : « le nombre de véhicules présents dans une section à un instant t+dt est égal au nombre de véhicules rentrés pendant dt moins le nombre de véhicules sortis pendant dt.

$$\underset{\text{véhicules présents à t+dt}}{C(x, t+dt)} = \underset{\text{véhicules présents en t}}{C(x, t)dt} + \underset{\text{véhicules entrants}}{Q(x, t)dt} - \underset{\text{véhicules sortants}}{Q(x+dx, t)dt}$$

En passant à la limite quand dt et dx tendent vers 0 on obtient

$$\frac{\partial C}{\partial t} + \frac{\partial Q}{\partial x} = 0$$

La deuxième équation, équation fondamentale, définit la vitesse comme le rapport du débit sur la concentration. En tout point de la route et à tout instant, il existe une relation liant la vitesse moyenne et la concentration.

$V = f(C)$

$V = V_f (1-C/C_{max})$

$Q(x,t)=C(x,t)*V(x,t)$

La troisième équation relie le débit à la concentration et peut être variable selon le modèle utilisé.

Soit le modèle est dit du premier ordre, et l'équation utilisée est une relation d'équilibre entre débit et densité, supposée vraie en tout point. Les transitions sont alors décrites comme des passages instantanés d'un état d'équilibre à un autre. Ces modèles ne représentent donc pas les états de transition.

Soit le modèle est dit du second ordre, et l'équation décrit alors l'accélération du flux. Les termes intervenant dans cette évolution, sont des termes de retard et d'anticipation. Ce modèle décrit plus finement les états de transition mais n'est pas utilisable directement (pas de solution analytique) mais uniquement en simulation.

Ces équations utilisent des paramètres comme la vitesse libre V_f (ou désirée) et la concentration critique $C_{max}/2$. Une fois définis et calibrés selon le type de réseaux, les paramètres sont considérés comme des caractéristiques du modèle et deviennent alors des constantes.

(c) Les modèles mésoscopiques

Ces modèles sont des intermédiaires entre la finesse des modèles microscopiques et la généralisation des modèles macroscopiques. Ils représentent le trafic sous forme de paquets de véhicules et traitent l'évolution de ces paquets individuellement. Par exemple, ils peuvent être une façon de décrire l'affectation du trafic, chaque paquet ayant une destination déterminée et un itinéraire donné.

(d) La simulation de l'écoulement

La simulation est la résolution numérique de ces modèles. Elle est fondée sur la discrétisation des variables de temps et d'espace. On calcule donc l'état du système à un instant t à partir de son instant à $t-\Delta t$, c'est la simulation dite « pas à pas ». Lorsque l'espace est discrétisé, certaines variables sont moyennées sur les zones comprises entre x et $x+\Delta x$ (par exemple, dans les modèles macroscopiques, la densité est supposée constante sur chaque section).

Le calibrage du modèle consiste à ajuster les paramètres du modèle. Un paramètre est une grandeur constante au cours de la simulation et est utilisé, soit pour décrire la physique du phénomène et la nature du réseau, soit pour prévenir des problèmes liés aux équations.

2) L'affectation du trafic

On cherche à définir la répartition des flux sur les différents itinéraires possibles compte tenu de la demande entre chaque origine et destination. Le principe retenu pour calculer l'affectation est que chaque usager cherche à minimiser un coût de déplacement, généralement défini par le temps de parcours.

Il existe deux conceptions différentes de calcul. Soit, en situation statique, la demande et l'offre sont supposées constantes au cours de la période étudiée. Le chemin de coût minimum

est obtenu en réalisant un équilibre où tous les itinéraires utilisés entre une origine et une destination ont un coût égal minimum. Soit, en situation dynamique, il existe une variation de la demande ou de l'offre au cours de la période étudiée.

(a) L'affectation prédictive statique

Cette affectation suppose une régularité quotidienne (la variation de la demande est identique d'un jour à l'autre). Les temps de parcours expérimentés un jour à une heure donnée fournissent une bonne prévision du temps de parcours le jour suivant à la même heure. Cette modélisation représente donc le comportement des usagers en situation stable.

(b) L'affectation réactive dynamique

Elle repose sur une définition du temps de parcours instantané. Elle correspond au comportement des usagers guidés en temps réel qui, à chaque point de choix, vont emprunter le chemin, considéré à cet instant comme le moins coûteux (en temps).

(c) La simulation pour l'affectation

On a souvent recours à une représentation individualisée des véhicules ou par paquet (regroupement des véhicules ayant la même destination et la même heure de départ). On utilise ensuite le temps de parcours pour calculer le plus court chemin.

3) Le problème des données

Les valeurs observées en trafic sont généralement très dispersées. Il y a deux explications à ces phénomènes. Tout d'abord il existe une erreur matérielle, due à l'éventuelle mauvaise qualité des capteurs. De plus, les phénomènes que l'on cherche à mesurer sont, par eux mêmes, bruités car il représente le comportement d'usagers difficilement mesurable.

Enfin dans le cas de l'affectation, les matrices origine-destination dynamiques nécessaires sont impossibles à obtenir. Sont alors utilisées des matrices statiques.

La méconnaissance des données initiales influe donc sur les résultats de la simulation. Il importe alors d'avoir une analyse critique des résultats par rapport aux données d'entrée. Ceci peut permettre de juger, qualitativement ou quantitativement, de la validité et de la portée des résultats.

II. Une procédure d'évaluation a priori des résultats issus d'un modèle de simulation

Dans ce paragraphe, on présente une démarche pour évaluer a priori les résultats issus de modèles de simulation. Celle-ci est basée sur la notion de prévision d'ensemble, concept très utilisé aujourd'hui en météorologie. Même s'il n'existe pas d'analogie directe entre la

météorologie et le trafic routier, les équations d'écoulement utilisées dans ces deux domaines se ressemblent car issues, en partie, de la mécanique des fluides. Nous avons également vu dans la présentation des modèles utilisés en trafic que le problème des données d'entrée justifie l'existence éventuelle d'une incertitude sur l'état futur. On va alors chercher à obtenir non pas un seul résultat mais plusieurs pour ainsi qualifier le résultat dit de contrôle c'est-à-dire le résultat ayant pour données d'entrée, les valeurs mesurées non perturbées.

1) Construction de l'ensemble de résultats

(a) Cas de simulation macroscopique

On considère un réseau global discrétisé en un certain nombre de tronçons ou sections. Comme on ne prend pas en compte les véhicules de manière individuelle, les trois variables que sont le débit, la vitesse et le taux d'occupation suffisent pour décrire l'écoulement du trafic.

Dans la construction d'un outil de simulation, il existe plusieurs étapes :

1) La première étape consiste à définir le modèle utilisé pour décrire l'écoulement du trafic par un certain nombre d'équations liant les variables.
2) Ensuite vient la phase de calibrage qui permet d'ajuster les paramètres du modèle aux caractéristiques physiques du réseau.

Ces deux étapes servent à construire le modèle.

3) Il convient de définir les conditions aux limites. Pour un modèle macroscopique les conditions aux limites sont les valeurs du débit, de la vitesse, du taux d'occupation au temps initial aux entrées et sorties du réseau et de chaque section.
4) La simulation du modèle à partir des conditions aux limites donne l'évolution du trafic dans le temps et l'espace.

En météorologie, afin de construire la prévision d'ensemble, les météorologues cherchent à intervenir dès la troisième étape. En effet, l'objectif n'est pas de remettre en cause le modèle (et ainsi d'intervenir dans la définition et le calibrage du modèle) mais plutôt de jouer sur les conditions aux limites afin d'envisager plusieurs possibilités futures.

Dans le cas de la simulation du trafic, on propose donc d'envisager un certain nombre de conditions aux limites représentant l'incertitude sur l'état initial (existant aussi en trafic).

Pour obtenir cet ensemble final deux méthodes sont possibles :

- soit en faisant varier les conditions limites
- soit en utilisant les résultats issus du même modèle et calculés sur un intervalle de temps antérieur au temps de la simulation (système du pauvre).

✓ Variation des conditions limites :

Il existe plusieurs possibilités pour faire varier les conditions limites.

- La plus simple est d'utiliser une méthode de Monte Carlo. On considère les valeurs mesurées des variables comme étant les moyennes respectives des distributions de probabilités représentant l'incertitude sur chaque variable. Les variations autour de cet état moyen peuvent être facilement générées en additionnant au hasard les nombres

caractéristiques des erreurs. Les valeurs aléatoires peuvent, par exemple, être des variations gaussiennes avec une moyenne nulle impliquant des erreurs dues aux mesures non biaisées.

- Une autre possibilité utilisée en météorologie est d'avoir recours aux vecteurs singuliers consistant à rechercher les régions de l'atmosphère les plus sensibles (celles où une petite erreur sur l'état initial est susceptible de croître extrêmement rapidement). Les axes de l'instabilité maximum peuvent être calculés à partir des vecteurs singuliers.

La première méthodologie peut être directement utilisée en trafic afin de faire varier les variables limites que ce soit le débit, la vitesse aux entrées et sorties du réseau. Les variables générées doivent être cohérentes avec les courbes fondamentales relatives à chaque entrée sortie. Soit on génère à la fois les variables débit et vitesse en vérifiant ensuite qu'elles sont en accord avec le diagramme fondamental. Soit on peut choisir de simuler uniquement le débit et associer à chaque valeur de débit une vitesse via la relation fondamentale. En ce qui concerne la deuxième méthodologie faisant référence aux vecteurs singuliers, un approfondissement s'avère nécessaire afin de voir si cette manière de perturber les conditions initiales peut être transposable au domaine du trafic routier.

Une fois, l'ensemble des conditions initiales généré, chaque point de cet ensemble sera le point de départ d'une simulation donnant ainsi un résultat. On obtient au final un ensemble de résultats qu'il conviendra d'analyser.

Cette méthode de variations des conditions initiales peut s'avérer coûteuse puisqu'elle demande des intégrations successives du modèle. Afin de palier à ce problème, une autre possibilité est d'utiliser le système du pauvre.

✓ Système du pauvre :

Dans cette manière de construire l'ensemble final, qu'une seule intégration du modèle est nécessaire. Cette construction utilise, en effet, les résultats déterministes issus du même modèle et calculés sur un intervalle de temps antérieur au temps de la simulation.

On suppose une variable scalaire x(t) associée à un modèle de simulation.
Hypothèse : on possède un échantillon de K résultats antérieurs provenant du même modèle $\{x_c(t_k), k \in [1\ldots K]\}$ ainsi que les observations correspondantes (que l'on appelle les valeurs vérifiées) $\{x_v(t_k), k \in [1\ldots K]\}$.
A l'instant t seul le résultat de la simulation contrôlé est disponible $x_c(t)$.

Méthode simple de construire un ensemble de résultats avec N membres :

Extraire du passé (à savoir de l'échantillon $\{x_c(t_k), k \in [1\ldots K]\}$) les N valeurs les plus proches de $x_c(t)$ et prendre comme résultat les observations correspondantes (attention : N<<K).

En notant I l'ensemble des N indices compris entre 1 et K correspondant aux données ainsi retenues l'ensemble de résultat de x(t) est $\{x_v(t_k), k \in I\}$ (auquel on rajoute le résultat de contrôle au temps t $x_c(t)$).
Cet ensemble peut être manipulé comme n'importe quel ensemble.

<u>Remarque 1 :</u>

Dans les deux méthodes de construction de l'ensemble final, on n'intervient pas au même moment. La première méthode de variations des conditions limites joue sur la troisième étape. Par contre, le système du pauvre ne demande en aucune manière de changer les

conditions aux limites mais envisage le passé déterministe en le comparant avec le résultat issu de la simulation.

Remarque 2 :

On a proposé dans ce paragraphe, deux méthodes de construction de l'ensemble finale, méthodes qui s'appliquent sur les variables tronçon par tronçon. De part la taille du réseau, cette méthodologie peut devenir lourde à mettre en place. Pour faciliter le travail, nous proposons éventuellement les mêmes démarches appliquées non plus aux trois variables mais à un ensemble d'indicateurs globaux.

Ces indicateurs peuvent être par exemple :

La longueur du tronçon i est représentée par l_i, son débit est q_i, sa vitesse v_i, longueur totale du trajet $L = \sum_{i=1}^{n} l_i$

Débit moyen $IQ = (\sum_{i=1}^{n} l_i \, q_i) / L$

Temps global passé $IT = \sum_{i=1}^{n} \frac{q_i \, l_i}{v_i}$

Vitesse moyenne $IV = \left[\sum_{i=1}^{n} q_i \, l_i \right] / \left[\sum_{i=1}^{n} \frac{q_i \, l_i}{v_i} \right]$

Temps de parcours moyen $TP = L / IV$

Certes, ces indicateurs ont un effet de lissage par rapport aux variables, et des indicateurs peuvent laisser penser que le trafic est fluide alors que l'étude variable par variable et tronçon par tronçon mettra en évidence des cas de congestion. Néanmoins, il peut s'avérer intéressant de regarder ces indicateurs. Appliquer la prévision d'ensemble sur ces indicateurs est moins fastidieux.

(b) Cas de simulation microscopique

Dans les cas de modèles macroscopiques, on justifie l'utilisation de la prévision d'ensemble par la présence d'erreurs de mesure et par le fait que les mesures sont par elles-mêmes bruitées. Dans le cas d'un modèle microscopique, le recours à la prévision d'ensemble peut s'expliquer, outre par les erreurs de mesure, par la présence du caractère stochastique de certaines variables.

Comme pour le cas macroscopique, il n'est pas question de changer les équations et paramètres du modèle.

Plusieurs variables sont stochastiques, notamment l'injection individuelle de véhicules fondée par exemple sur une loi de Poisson (voir Annexe) ayant pour paramètre l'inverse du débit. L'idée la plus simple est d'envisager de générer plusieurs fois la même loi ce qui peut donner des temps d'arrivée différents.

On peut également, comme pour le cas macroscopique, faire varier le débit et la vitesse initiaux aléatoirement dans les limites que l'on connaît (notamment par rapport à l'erreur de mesure).

Si on fait varier le débit, on peut générer une seule fois la loi de Poisson pour chaque valeur de débit obtenue. On peut aussi penser à générer pour chaque valeur du paramètre plusieurs fois la loi d'inter arrivée. Pour ce qui est des vitesses, la vitesse du premier véhicule est connue puis ensuite on prend une fonction de la distance du véhicule entré précédemment dans le réseau. Il est donc possible d'obtenir plusieurs résultats pour la même échéance et ainsi de construire un ensemble de résultats.

2) Analyse statistique de l'ensemble final

Quelle que soit la méthode retenue, on obtient alors non plus un seul résultat mais un ensemble qu'il est nécessaire d'analyser. Nous proposons dans ce paragraphe quelques pistes statistiques pour analyser cet ensemble. Finalement, cet ensemble contient des vecteurs à trois dimensions donnant pour chaque élément le débit, la vitesse et le taux d'occupation. Les trois variables ne sont pas indépendantes.

En premier lieu, même si les trois variables ne sont pas indépendantes (de part la construction du modèle, elles doivent vérifier certaines équations), on peut faire une étude unidimensionnelle. Pour chaque variable on trace l'histogramme associé à l'ensemble. La dispersion de l'histogramme peut mettre en évidence des cas d'aberration (par exemple, si avec la vitesse et le débit, on a dans le même ensemble des cas de congestion et de fluidité pour le même temps). De plus, on peut analyser les valeurs de la moyenne, de l'écart type et du coefficient de variation. Il est nécessaire d'être prudent en fonction de la taille de l'ensemble. Si on traite l'ensemble sur tout l'axe celui-ci sera assez important pour que les statistiques soient robustes. Par contre, si l'on préfère travailler par section, les ensembles obtenus peuvent avoir une taille trop petite. Les statistiques peuvent alors être rendues plus robustes par une méthode du bootstrap. Il sera également possible de donner des seuils critiques empiriques voire même de faire une estimation non paramétrique de la densité pour chaque variable.

Cette analyse des différents histogrammes donne directement une première idée de la confiance à accorder au résultat de contrôle. Dans le cas de grande sensibilité aux données d'entrée, cette étude peut mettre en évidence une grande dispersion dans les résultats provenant d'une petite variation au départ. Si en changeant un peu les conditions aux limites, on aboutit à un ensemble de résultats mettant en évidence des résultats très diverses, on peut penser que le résultat issu de la simulation n'est pas représentatif. Au contraire, si l'ensemble est très homogène, la moyenne sera alors une des caractéristiques les plus probables.

Les variables étant liées les unes aux autres on ne peut s'en tenir à un regard unidimensionnel. Il est souhaitable de se placer aussi dans un espace à trois dimensions. Dans cet espace, il convient de regarder comment les résultats se ressemblent ou diffèrent des uns des autres. On peut donc essayer d'envisager une classification sur les résultats. Comme en météorologie, cette classification doit mettre en évidence les profils se ressemblant mais aussi donner un poids aux situations extrêmes (si elles existent). Le nombre de situations divergentes de la situation envisagée donnera alors une idée de la qualité du résultat de contrôle. Pour cette classification, il conviendra de trouver un critère d'agrégation ainsi qu'une distance adaptés au problème. En effet, les trois variables ne s'expriment pas avec les

mêmes unités. Il convient alors de travailler sur les variables centrées normées de manière à supprimer l'effet de taille. En analysant les variables normalisées, une possibilité est de faire une classification avec pour distance, la distance euclidienne, et pour critère d'agrégation, celui de la perte d'inertie minimale. D'autres possibilités sont aussi envisageables.

Une analyse des données considérant l'ensemble final comme un tableau d'individus ayant chacun trois caractéristiques peut donc s'avérer un bon moyen pour qualifier le résultat de contrôle. Le meilleur cas sera donc celui où une seule classe homogène existe.

Enfin, on peut également travailler dans des hyperplans où l'une des trois variables est constante ou appartenant à un intervalle. Dans chaque hyperplan, on trace le nuage de points. On compare alors la forme obtenue à la forme attendue calculée par exemple avec l'historique

Conclusion générale

Le concept d'évaluation à priori d'une prévision est fortement utilisé aujourd'hui en météorologie. Au cours de ce travail il s'agissait de comprendre pourquoi et comment les méthodes de prévisions en météorologie nécessitaient un calcul de l'incertitude et non pas d'étudier entièrement les techniques très complexes de prévision du temps. Les modèles numériques décrivant l'évolution des paramètres sont un moyen comme le sont les observations pour prévoir le temps futur. Cette méthode numérique ne nécessite en aucun cas le recours à la statistique. Par contre, cette dernière intervient lorsque l'on cherche à obtenir une prévision d'un phénomène météorologique, celui-ci étant une combinaison linéaire des paramètres prévus numériquement. Un des problèmes majeurs de ces équations d'évolution est leur sensibilité aux données initiales. Or, il semble parfaitement impossible de connaître l'état de l'atmosphère en tout point et de manière exacte. Cette notion d'incertitude entraîne alors une grande variabilité dans les prévisions issues de cet état initial. L'objectif des chercheurs devient alors d'étendre la limite de validité des prévisions tout en améliorant le niveau moyen de fiabilité de ces prévisions.

Les méthodes d'interprétation des prévision numériques ont mené à l'introduction d'un indice de confiance dans les prévisions. Pour connaître à l'avance l'incertitude d'une prévision numérique, il faudrait connaître l'impact des différentes sources d'erreurs. En considérant la méconnaissance de l'état initial comme la cause d'erreur la plus importante, il suffit donc de perturber légèrement cet état (dans les limites de ce que l'on sait) pour produire une nouvelle prévision. On aboutit alors à la notion de prévision d'ensemble où non plus une seule valeur est prévue mais, par exemple une cinquantaine. Une approche consiste alors à ne voir dans le résultat de la prévision d'ensemble que les distributions de probabilités des différentes variables météorologiques. Ainsi la probabilité de l'événement « beau temps chaud » défini par « pas de précipitations, température supérieure à 25 degrés et vent inférieur à 5m/s », est donnée immédiatement par le nombre de prévisions pour lesquelles ces différentes occurrences sont prévues simultanément, rapporté au nombre total de prévisions.

La prévision d'ensemble fournit donc au prévisionniste un ensemble de techniques adaptées à la prévision du temps. Celles-ci permettent d'évaluer à priori l'incertitude de la prévision. Une classification des prévisions de l'ensemble met en évidence l'évolution du temps la plus vraisemblable, ainsi que les alternatives possibles à cette évolution. Un indice de confiance peut quantifier l'incertitude indiquée par le nombre et l'importance de ces alternatives. La prévision d'ensemble ouvre également la voie à une formulation probabiliste de la prévision du temps. La probabilité de tout événement météorologique quantifiable peut être facilement calculée à partir de la distribution des prévisions de l'ensemble.

Certes, il n'existe pas d'analogie directe entre la météorologie et le trafic. Par contre, l'approche de calcul de l'incertitude utilisée en météorologie peut être transposable au trafic. C'est dans cette optique que nous avons travaillé sur le modèle de prévision du dispositif Bison Futé permettant de prévoir le trafic journalier un an à l'avance. L'objectif de cette application était donc d'utiliser certains concepts d'évaluation des prévisions (mentionnés ci-dessus) afin de qualifier à priori la prévision du trafic issue du modèle. Nous avons d'abord présenté la méthode de prévisions et procédé à une analyse à posteriori des résultats. La qualité des résultats obtenus est bonne mais dépend néanmoins du type de jours. Il reste en effet des erreurs significatives lorsque le jour de prévision est un jour dit exceptionnel. Pour ces types de jours, le débit est très important et donc l'enjeu d'avoir une bonne prévision est capital.

Pour chaque jour, un ensemble de prévisions a été construit via le concept de système du pauvre appliqué au débit journalier relatif. L'analyse de cet ensemble (valeurs statistiques, dispersion…) a permis alors d'associer à chaque prévision dite de contrôle un indice de

confiance. Il convient de noter que cet indice n'est en aucun cas une probabilité mais un qualificatif de la prévision variant de un à quatre. De plus, la prévision d'ensemble permet aussi de tirer à priori d'autres renseignements quantitatifs sur le débit relatif.

Ainsi, dans les cas où l'indice de confiance associé était faible, il fut à chaque fois possible de donner un seuil critique inférieur. Dans ces configurations, on ne pouvait donc pas donner une valeur exacte pour le taux mais un seuil pour lequel on est sûr que le taux réel sera supérieur. De même on a donné un seuil critique supérieur donnant une valeur de taux à ne pas dépasser. De plus, l'analyse de la dispersion de l'ensemble a souvent permis de mettre en évidence des cas de sous estimation ou de sur estimation. Même si, dans ce cas, le qualificatif associé à la prévision n'est pas bon, il est toutefois possible de conclure à priori si le débit réel sera supérieur ou inférieur à ce que l'on prévoit. On peut alors envisager un rectificatif de la prévision.

Enfin, il existe de nombreux jours où la confiance en la prévision est bonne voire excellente.

Avant de clore cette conclusion, il est nécessaire de faire plusieurs remarques. Certes, le modèle est toujours améliorable mais l'objectif ici était bien de faire une évaluation à priori des prévisions issues de ce modèle.

Ensuite dans tout le travail effectué on n'a pas utilisé la dimension spatiale du trafic puisque nous avons travaillé sur un point précis de l'espace (St Arnoult), et sur une valeur agrégée du trafic (débit journalier). De plus, l'horizon de prévisions étant d'un an, il est impossible de tenir compte de l'évolution dynamique du trafic. Dans la construction de l'ensemble de prévisions, la première sélection porte sur un critère qualitatif à l'aide des variables calendaires. Or, les codifications calendaires sont fixées indépendamment du sens de circulation. Certains jours sont donc répertoriés comme exceptionnels or le débit dans le sens des départs est celui d'un jour ordinaire. La prévision d'ensemble permet en partie de palier à ce problème puisque l'on détecte alors une sur estimation.

Enfin, pour ce genre d'évaluation à priori, il faut d'abord avoir une très grande confiance dans le modèle de prévisions. En effet, l'erreur doit essentiellement provenir de l'effet de variables exogènes non prévisibles et non pas d'un mauvais calibrage du modèle. De plus, c'est le même modèle qui doit servir à la prévision et à l'estimation sur l'historique. Il semble alors très délicat et difficile d'appliquer ce genre de concept en temps réel où l'équation de prévision (type ARMA) est amenée à être modifiée (ne serait-ce que dans les valeurs des paramètres).

Nous proposons d'appliquer ces méthodes dans le cas de la simulation de trafic. Nous avons tenté de montrer que cette approche pouvait être applicable pour l'évaluation a priori des résultats issus des outils de simulation. Certains phénomènes peuvent rendre incertain l'écoulement du trafic (erreur sur les mesures, présence de variables stochastiques…). Le recours au système du pauvre ou à la variation des données initiales permet alors l'obtention d'un ensemble de résultat et non plus d'un seul résultat. L'analyse statistique de l'ensemble donne une idée a priori de la qualité du résultat. Ces méthodologies permettent donc de qualifier le résultat sans pour autant remettre en question le modèle ou la technique utilisés.

Certes un modèle de prévision est toujours améliorable, mais avoir recours à la prévision d'ensemble, ce n'est pas modifier le modèle mais bien, de faire une évaluation a priori du résultat issu de ce même modèle. C'est bien dans cet objectif que nous l'avons utilisé et que nous proposons de l'utiliser dans le cas de la simulation. Finalement cela revient à adopter une attitude critique par rapport aux résultats dits numériques afin de pouvoir les qualifier.

Bibliographie

Atger F., 1999 : The Skill of Ensemble Prediction Systems, *Mon. Wea. Rev.*,127, 1941-1953.

Atger F., 2000 : La prévision du temps à moyenne échéance en France, *La Météorologie* 8ème série, 30, 61-86.

Calot G., 1971 : Cours de calcul des probabilités ; *Dunod*

Certu, 2000 : Simulation dynamique du trafic routier ; *Collections du Certu dossier 106*

Coiffier J., 2000 : Un demi-siècle de prévision numérique du temps, *La Météorologie* 8ème série, 30,11-31.

Couton F. , Danech-Pajouh M. , Debauvais R. (Sept. 1996) Les modèles de prévision et le dispositif Bison Futé, *Convention INRETS-DSCR.*

Dupuis P., 1999a : La prévision à moyenne échéance. *Met. Mar.*, 183, 3-8.

Dupuis P., 1999b : Prévision d'ensemble et indice de confiance. *Met. Mar.*, 183, 9-12.

Dynamic meteorology : data assimilation methods :1981, *Bengtsson, Lennart. DYN Ed. ;Ghil, Michael. Ed. ; Kallen, Erland. Ed. - New York NY ; 81 Heidelberg ; Berlin : Springer-Verlag,* 330 p. (Applied mathematical sciences ; 36).

Efron B., Tibshirani R.J., 1993 : An introduction to the Bootstrap, *Monographs on statistics and Applied Probability n° 57*, Chapman & Hall

Le Dimet F. X., Talagrand O., 1986 : Variational algorithms for analysis ans assimilation of meteorological observations : theorical aspects, *Tellus*, 38A, 97-110.

Leith C.E., 1974: Theoritical skill of Monte-Carlo forecasts, *Mon. Wea. Rev.*, 102, 409-418.

Molteni F, Buizza R, Palmer T.N., 1990 : The EMCWF ensemble prediction system : methodology and validation, *Q.J.R. Meteorol. Soc.*, 122, 73-119.

Palmer T. N., 2000, Predicting uncertainty in forecasts of weather and climate, *Rep. Prog. Phys.*, 63, 71-116.

Rochas M., Javelle J-P, 1993 : La météorologie, *collection Comprendre, Syros/Alternatives*, 262 p.

Roux F. , 1993 : Le temps qu'il fait, *Documents Payot, Editions Payot & Rivages*, 315 p.

Talagrand O., Courtier P., 1987 : Variational assimilation of meteorological observations with the adjoint vorticity equation. I : Theory, *Q.J.R. Meteorol. Soc.*, 113, 1311-1328.

Talagrand O., Courtier P., 1987 : Variational assimilation of meteorological observations with the adjoint vorticity equation. II : Numerical results, *Q.J.R. Meteorol. Soc.*, 113, 1329-1347.

Talagrand O., Vautard R., Strauss B, 1999 : Evaluation of probabilistic prediction systems, *Compte-rendu Atelier Prédictabilité ECMWF*.

Toth, Z., and E. Kalnay, 1997, Ensemble Forecasting at NCEP and the Breeding Method, *Mon. Wea. Rev.*, 125, 3297-3319.

Wilks D., 1995 : Statistical methods in the Atmospheric sciences, *Academic Press, INC.*

Ziani A., Danech-Pajouh M., (Nov. 1998), Prévision du trafic journalier, Modèle linéaire généralisé (spécification pour l'implémentation informatique).

Annexes

Annexe : généralités de la théorie du trafic

Dans la modélisation du trafic, deux types de modèles d'écoulement du trafic sont distingués :

- Les modèles Microscopiques
- Les modèles Macroscopiques

Les modèles microscopiques analysent le comportement individuel des automobilistes qui sera alors traduit par des lois de poursuite. La simulation microscopique fournit une grande richesse d'informations, mais un très grand nombre de données à recueillir et un temps de calcul trop élevé font que cette simulation si fine, devient trop lourde pour les objectifs de la régulation.

Les modèles macroscopiques ignorent les véhicules individuels et n'étudient que le comportement de leur flot. La simulation macroscopique a l'avantage de faire intervenir un nombre peu élevé de variables et de leurs données requises et présente donc un calcul simple pour un temps moindre. Son principal inconvénient est une faible adaptation au trafic urbain.
Pour des propos de régulation le modèle utilisé est un modèle macroscopique.

I. Variables Macroscopiques du Trafic :

La description macroscopique de l'écoulement du trafic utilise des variables exprimant le comportement moyen des flots de véhicules sur une section de route donnée.
Les variables utilisées sont :

- La densité ρ (nombre de véhicules par unité de longueur)
- Le débit q (nombre de véhicules par unité de temps)
- La vitesse v du flot

- Densité :

Soit $N(x_1, x_2, t)$ le nombre de véhicules se trouvant à l'instant t sur le segment $[x_1, x_2]$. La densité moyenne $\overline{\rho}(x_1, x_2, t)$:

$$\rho(x_1, x_2, t) = \frac{N(x_1, x_2, t)}{x_1 - x_2} \qquad (1)$$

Généralement cette densité est estimée à partir d'une variable mesurable, le taux d'occupation, noté *t.o*. Cette variable correspond au temps durant lequel le capteur a été occupé. Le taux d'occupation est directement lié à la densité par la relation (2) suivante :

$$t.o = (L + l)\overline{\rho} \qquad (2)$$

où L est la longueur moyenne d'un véhicule et l celle d'un capteur.

Sous l'hypothèse de continuité du flot, on appelle densité au point x à l'instant t à la fonction définie par :

$$\rho(x, t) = \lim_{\Delta x \to 0} \overline{\rho}\left(x - \frac{\Delta x}{2},\ x + \frac{\Delta x}{2},\ t\right) \qquad (3)$$

Et pour une valeur petite de Δx :

$$\rho(x,t) \cong \bar{\rho}\left(x-\frac{\Delta x}{2},\ x+\frac{\Delta x}{2},\ t\right) \quad (4)$$

- **Débit :**

Soit $N(x, t_1, t_2)$ le nombre de véhicules passés en x entre les instants t_1 et t_2. On appelle débit moyen la grandeur $q(x_1, x_2, t)$ vérifiant :

$$\bar{q}(x,t_1,t_2) = \frac{N(x,t_1,t_2)}{t_1 - t_2} \quad (5)$$

Sous l'hypothèse de continuité du flot, on appelle débit au point x à l'instant t, à la fonction définie par :

$$q(x,t) = \lim_{\Delta t \to 0} \bar{q}\left(x,\ t-\frac{\Delta t}{2},\ t+\frac{\Delta t}{2}\right) \quad (6)$$

Et pour une valeur petite de Δt :

$$q(x,t) \cong \bar{q}\left(x,\ t-\frac{\Delta t}{2},\ t+\frac{\Delta t}{2}\right) \quad (7)$$

- **Vitesse :**

On appelle vitesse moyenne d'un flot de véhicules au point x à l'instant t, la fonction v vérifiant :

$$v(x,t) = \frac{q(x,t)}{\rho(x,t)} \quad (8)$$

II. Lois fondamentales

L'écoulement du trafic macroscopique repose sur trois principes de base :

- La représentation continue.
- La loi de la conservation de la masse.
- L'hypothèse du diagramme fondamental.

- **Représentation continue**

L'écoulement du flot le long d'axe routier présente des analogies avec la théorie hydrodynamique des fluides. Ce type de représentation interprète un flot de véhicules comme étant un milieu continu.

Sous l'hypothèse de continuité, il nous faudra considérer d'une part qu'un véhicule peut être ramené à un point (en négligeant sa dimension) et d'autre part que n'importe quelle section d'une chaussée contient un nombre déterminé de véhicules aussi petite soit elle.

Il est très difficile de cerner le type d'approximation que réalise ce premier principe. Intuitivement, il s'applique lorsque la circulation est dense.

- **Loi de Conservation**

Le second principe de cette théorie hydrodynamique est la conservation de la masse. La variation du nombre de véhicules se trouvant sur une section de route $[x_1, x_2]$ doit être égale à la différence entre le nombre de véhicules passant en x_1 durant l'intervalle de temps $[t_1, t_2]$ et le nombre de véhicules passant en x_2 pendant le même intervalle de temps.

En effet :

$$\int_{x_1}^{x_2} [\rho(x,t_2) - \rho(x,t_1)]dx = \int_{t_1}^{t_2} [q(x_2,t) - q(x_1,t)]dt \quad (9)$$

Ce principe est très adéquat puisque d'une part il correspond à la réalité du trafic et que d'autre part, il n'entre pas en conflit avec la théorie discrète de l'écoulement, l'équation (9) pouvant s'écrire :

$$N(x_1, x_2, t_2) - N(x_1, x_2, t_1) = N(x_2, t_1, t_2) - N(x_1, t_1, t_2) \quad (10)$$

- **Relation Vitesse -Densité :**

La relation vitesse -densité décrit le comportement des usagers. Cette relation exprime la vitesse comme une fonction de la densité :

$$v(x,t) = v(\rho(x,t)) \quad (11)$$

Pour une analyse des conditions du trafic, on remarque que lorsque la route est vide (la densité est nulle), un véhicule circule avec la vitesse désirée par le conducteur, appelé aussi vitesse livre :

$$v(\rho = 0) = v_f \quad (12)$$

En cas contraire, quand le trafic est saturé ($\rho=\rho_{max.}$) la vitesse est nulle :

$$v(\rho = \rho_{\max}) = 0 \quad (13)$$

Une équation F dite fonction comportementale illustre bien la liaison entre l'équation (11) et la représentation du comportement des automobilistes.
Une formule généralisée représentant ce comportement satisfaisant aux conditions limites est :

$$F(\rho) = v_d \left[1 - \left(\frac{\rho}{\rho_{\max}} \right)^l \right]^p \quad (14)$$

où $p \geq 1$ et $l > 0$ sont des paramètres réels.

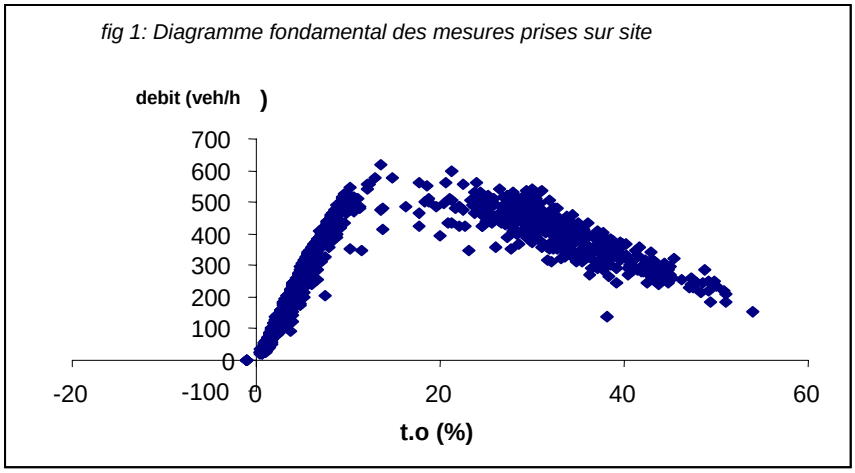
III. Diagramme Fondamental :

Substituons $F(\rho)$ dans l'équation (8) pour obtenir la relation débit-densité :

$$q = Q(\rho) = \rho\, F(\rho) \quad (15)$$

Cette relation est connue sous le nom de *diagramme fondamental*.
Les formules modélisant le diagramme fondamental sont nombreuses à cause de la forte dispersion de données réelles (fig 1)

fig 1: Diagramme fondamental des mesures prises sur site



En général le diagramme fondamental se présente sous la forme décrite par la figure suivante :

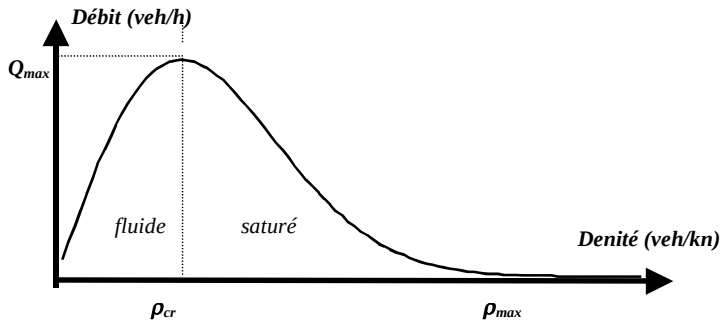


fig 2 : Diagramme fondamental

On appelle densité critique ρ_{cr} la densité pour laquelle le débit atteint son maximum Q_{max}. C'est la densité optimale pour la circulation, elle correspond à l'utilisation totale de la capacité de l'infrastructure.
L'état de la circulation est fluide dans l'intervalle [0, ρ_{max}], c'est à dire, le débit augmente avec la densité. Au voisinage de la densité critique, le débit est instable et au-delà de ce seuil critique, si la densité augmente le débit diminue et l'évolution du trafic tend vers la congestion. Près de la densité maximale le régime atteint la situation de blocage

Annexe : Construction de l'intervalle de confiance de la prévision issue du modèle GLM

Pour cela avec les notations utilisée dans la partie présentant l'analyse de variance on peut calculer $(X^{'}X)^{-1}$ où X est le tableau disjonctif des variables explicatives de l'historique.

La matrice de variance covariance sur l'historique est alors $\overset{\wedge}{\sigma}^{2} X'(X^{'}X)^{-1}X$ avec $\overset{\wedge}{\sigma}^{2}=\frac{\|y-X\beta\|^{2}}{n-p}$.

Pour chaque prévision de l'année 1998 on peut obtenir l'écart type σ_{y_p} qui sera donc la racine des termes diagonaux de la matrice $\overset{\wedge}{\sigma}^{2} X'_p (X^{'}X)^{-1}X_p$ où X_p est la matrice des variables explicatives pour l'année 98.

L'intervalle de confiance de la valeur prédite y peut s'obtenir par
$y \pm \sigma_{y_p} t(1-\alpha/2)$ avec n-p degrés de liberté pour Student.

Annexe : Test d'échantillons

L'une des premières choses à faire est de vérifier que l'ensemble de prévisions obtenu constitue un échantillon d'une variable aléatoire c'est-à-dire que les X_i sont indépendants, identiquement distribués et de même loi. Pour cela, on est amené à utiliser des tests non paramétriques Pour cela on pourra utiliser le test des runs up down (attention c'est un test asymptotique donc à utiliser avec beaucoup de prudence).

H0 : les observations forment un échantillon

H1 : alternative

On considère le signe de l'expression $X_t - X_{t-1}$ .Si cette quantité est positive on lui associe la valeur 1, sinon la valeur0.

Soit R la variable aléatoire représentant le nombre total des suites de 0 et 1. Sous H0 on a

$$E(R) = 2n - \frac{1}{3}$$

$$Var(R) = 16n - \frac{29}{90}$$

La loi de R (quand n est assez grand) est bien approximée par une loi normale.

La région de rejet est donc défini par $\frac{|E - E(R)| + 1/2}{\sigma_R} > q_{1-\alpha/2}$ où $q_{1-\alpha/2}$ est le quantile $1-\alpha/2$ d'une normale centrée réduite.

On peut également effectuer le test à la médiane qui est le même sauf que R est le nombre de suites de différences à la médiane (on code 1 ou 0 selon que la valeur est au-dessus et en – dessous de la médiane).

Annexe : Méthodes non paramétriques

Les modèles statistiques non paramétriques ne peuvent être indexés que par un paramètre évoluant dans un espace vectoriel à une infinité de dimensions c'est-à-dire un espace fonctionnel. Un exemple classique de modèle non paramétrique pour une variable aléatoire réelle est obtenu en postulant que la loi de probabilité de cette variable est absolument continue et possède une densité uniformément continue ou encore le densité de probabilité est lipschitzienne sur $\Re$.

I. Tests d' adéquation de lois

✓ Adéquation graphique à une famille de lois : méthode du QQ-plot :

Cette méthode est un test graphique d'adéquation à une famille de lois.

Nous testons (Ho) : l'échantillon est issu d'une famille de loi
$\{F_{\mu,\sigma}, \mu \in R, \sigma>0, F_{\mu,\sigma}(x) = F_0((x-\mu)/\sigma)\}$, où F_0 est une fonction de répartition continue, contre sa contraposée (H_1).

Pour cela, nous traçons le graphe des points ($F_0^{-1}(S_k)$, $X_{(k)}$), où $X_{(k)}$ est la valeur de la k^{ieme} observation dans l'échantillon ordonné, F_0 la fonction de répartition d'une loi de la famille testée et S_k le pas du test, égal à k/n ou k/(n+1) .

Si (Ho) est vraie, alors les points du graphe sont alignés sur une droite. De plus, l'équation de la droite permet d'estimer les valeurs μ et σ : la droite coupe l'abscisse en $-\mu/\sigma$ et a pour pente $1/\sigma$.

Nous pouvons tracer sous le logiciel SAS Insight les QQ-plots pour les lois normales, lognormales, exponentielles et de Weibull (respectivement notées N, L, E et W).

✓ Test d'adéquation de Kolmogorov :

Rappelons le principe de ce test :

Soit $X_1, X_2, \ldots, X_n$ un échantillon d'une loi F et soit F_0 une loi continue.
On veut tester :

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

Pour cela on utilise la distance entre la fonction de répartition empirique et la fonction de répartition théorique.

La statistique de Kolmogorov-Lilliefors est définie par :

$$D_n = \sup_x \left| F_n(x) - F_0(x) \right|$$

qui peut aussi s'écrire :

$$D_n = \max\left\{D_n^+, D_n^-\right\}$$

avec $D_n^+ = \max\left(\frac{i}{N} - F_0\left(X_{(i)}\right), 0\right)$ et $D_n^- = \max\left(F_0\left(X_{(i)}\right) - \frac{i-1}{N}, 0\right)$

où $X_{(i)}$ est la $i^{ème}$ observation de l'échantillon, et N le nombre total d'observations.

Le test au seuil α , associé à cette statistique est défini par la région critique de la forme :

$$\left\{D_n \geq c_\alpha\right\}$$

où c_α est le quantile $(1-\alpha)$ de la table de Kolmogorov-Lilliefors.
On utilise la statistique de Kolmogorov-Lilliefors au lieu de la statistique de Kolmogorov-Smirnov dans la mesure où on travaille avec les paramètres estimés.

Ce test a été effectué grâce au logiciel SAS Insight.

II. Estimation non paramétrique de la densité

Pour estimer des densités des lois de séries de données, il est donc possible d'avoir recours à des méthodes non paramétriques. Dans ces méthodologies, les densités ne sont plus caractérisées par des paramètres mais par des familles de fonctions.

La fonction de répartition empirique joue un rôle crucial dans l'étude de la loi d'un échantillon. Cependant elle ne permet pas l'obtention de résultats très précis sur la structure de cette loi. En revanche, lorsque les variables de l'échantillon admettent une densité f celle-ci donne beaucoup plus d'informations sur la loi (dispersion, modes…). L'estimation de f est donc un problème fondamentale de la statistique non paramétrique.

Au départ, on peut estimer la densité en un point en comptant le nombre d'observations appartenant au voisinage de ce point (intervalle de la forme $\left[\, x - \frac{h_n}{2} ; x + \frac{h_n}{2} \,\right]$ sur R, avec h_n réel strictement positif)).
On aura alors comme estimateur de f

$$f_n(x) = \frac{F_n(x + h_n/2) - F_n(x - h_n/2)}{h_n}$$ où F_n est la fonction de répartition empirique.

Une des méthodes les plus utilisées et présentée ici est la méthode des estimateurs à noyaux ; cette méthode n'est valable que si les séries forment un échantillon, c'est-à-dire, si les données sont indépendantes et identiquement distribuées. Pour tester si les séries forment un échantillon on peut avoir recours au test présenté dans l'annexe suivante.

Un noyau est une fonction K sur les réels bornée telle que $\int_{-\infty}^{+\infty} K(u).du = 1$.

On dit qu'un noyau sur R est de Parzen-Rosenblatt si :

$$\lim_{\|x\| \to \infty} (\|x\| K(x)) = 0$$

L'estimateur de Parzen-Rosenblatt de la densité f associé au noyau K est de la forme :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h} \right)$$

où h est le bandwidth de l'estimation dépendant de n. Nous supposerons que le noyau K est positif ce qui permet d'avoir un estimateur de f qui soit aussi une densité de probabilité.

On utilisera le logiciel SAS pour faire ces estimations, et le module Insight.
Le logiciel SAS Insight propose trois noyaux différents :

- le noyau triangulaire $K(u) = (1-u)_+$
- le noyau normal $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
- le noyau d'Epanechnikov $K(u) = \frac{3}{4}(1-u^2)_+$ (appelé quadratic par le logiciel SAS Insight)

L'influence du noyau sur la qualité de l'estimation est négligeable en comparaison avec l'influence du bandwidth.

h va déterminer la qualité de l'estimation :

- Si h est trop élevé, le biais de l'estimation sera trop important ;
- Si h est trop faible, la variance de l'estimation sera trop importante.

Il s'agit donc de trouver h de façon à obtenir un équilibre biais-variance.
h ne peut pas être calculé, mais il peut être estimé : le logiciel SAS Insight estime h en minimisant la statistique AMISE, qui est une approximation de la moyenne intégrée des carrés des erreurs.
On a la formule suivante :

$$\mathrm{MISE}(f) = \int E[f_n(x) - f(x)]^2 dx = \int (b^2(x) + \sigma^2(x)) dx$$

On remarque que le choix du noyau influe peu sur la valeur de cette erreur asymptotique (quand h tend vers zéro et nh vers l'infini) par contre le choix de h est crucial.

Annexe : LOI DE POISSON

I. Définition et présentation

Une approche pour présenter la loi de Poisson est de considérer le cas limite de la loi binomiale quand n (taille de l'échantillon) est grand (n>50) et θ (paramètre de la loi) est petit (θ<0.1).
Pour la loi binomiale on a E(x)=n*θ.
Pour la loi de Poisson (en partant de la binomiale) on trouve donc quand n est grand

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0,1,....$$

Alors E(X)=λ et V(X)=λ

✓ Processus de Poisson (deuxième manière d'introduire la loi de Poisson)

Cet autre mode de génération permet de considérer la loi de Poisson non plus comme une loi limite mais comme une loi exacte.

On suppose qu'un seul événement arrive à la fois, que le nombre d'événements se produisant pendant une période T ne dépend que de la durée de cette période et que les événements sont indépendants. Soit E(N)=cT où c est la cadence.

Si le nombre moyen d'événements par unité de temps est c, on démontre que la probabilité d'obtenir n événements pendant T est

$$P(N = n) = e^{-(cT)} \frac{(cT)^n}{n!}$$

On dit qu'une suite indéfinie d'événements A_1, A_2, …., A_i,… se réalisant aux dates aléatoires T_1, T_2, …., T_i,…obéit à un processus de Poisson si les hypothèses suivantes sont satisfaites :

(a) Le nombre d'événements A qui apparaissent entre les dates t et t+h est indépendant du nombre d'événements arrivant entre les dates 0 et t ;
(b) La probabilité qu'un événement A *au moins* apparaisse entre t et t+h est égale à ph+o(h) où p est une constante par rapport à t et où o(h) est un infiniment petit d'ordre supérieur à h lorsque h tend vers zéro [$(1/h)o(h) \to 0$] ;
(c) La probabilité que deux événements A ou plus apparaissent entre t et t+h est de l'ordre de o(h).

On peut prendre, comme exemple de processus de Poisson, l'arrivée de véhicules en un point donné d'une route.

L'hypothèse (a) est satisfaite si les arrivées de véhicules sont indépendantes (il n'y a pas de feux de circulation dans le voisinage, les véhicules n'ont aucune difficulté à se doubler).

On peut imaginer ainsi que l'**on observe les arrivées de véhicules sur une autoroute**, loin d'une agglomération, dans une région non accidentée.

L'hypothèse (b) est satisfaite si, au cours de la période d'observation au moins, le rythme d'arrivée des véhicules est constant (mesuré par p) : **la période d'observation est homogène**. Cette hypothèse revient à supposer que l'on ne prolonge pas la période d'observation outre mesure car la circulation automobile connaît évidemment un mouvement saisonnier diurne.

La troisième hypothèse exprime **qu'au cours d'un intervalle de temps infinitésimal, la probabilité d'arrivée de deux véhicules est très petite** par rapport à la probabilité d'arrivée d'un seul véhicule qui, elle-même, est petite.

II. Génération de la loi de Poisson (variable aléatoire discrète) :

✓ Algorithme

- s=0, n=0
- tant que s< λ on répète
  - U suit une loi uniforme [0,1]
  - s = s – lnU
  - n=n+1
- X=n-1

✓ Autre manière de présenter la génération de loi de Poisson

Générer U_1, U_2, U_3,…., U_n loi uniforme sur [0, 1]
Et faire le produit jusqu'à ce que

$\prod_{i=1}^{X+1} U_i < e^{-\lambda}$ c'est-à-dire

$-\ln(\prod_{i=1}^{X+1} U_i) > \lambda \Rightarrow$ délivrer X qui suit une loi de Poisson

En ce qui concerne la simulation du trafic routier dans le cas de modèles microscopiques, l'injection individuelle des véhicules dans le réseau suit une loi de Poisson de paramètre l'inverse du débit. Comme on travaille en secondes, il faut multiplier cette valeur par 3600. On aura ainsi une valeur de paramètre faisant référence à des secondes et on génère alors des temps d'arrivée en seconde. On vérifiera ensuite si en générant l'arrivée de, par exemple, 800 véhicules par heure, la somme des temps d'arrivée trouvés vaut bien environ 3600 secondes.

III. Utilisation de la loi exponentielle négative

L'arrivée des véhicules est générée aux points d'origine, c'est-à-dire à la périphérie du réseau étudié. Cette génération se fait en accord avec une loi de distribution basée sur le volume de véhicules.

Par exemple pour la loi exponentielle négative on utilise l'expression :

$$h = (H - h_{min})[-\ln(1-R)] + H - h_{min}$$

h= temps séparant l'arrivée de deux véhicules
H= temps moyen (3600/q)
h_{min}= temps minimum (ex : 1.2 sec/veh)
R= nombre aléatoire issu d'une loi uniforme [0,1]

Pour garantir que le débit exact sera ainsi simulé pendant le temps désiré (ex : 15 min) le modèle peut calculer le facteur K :

$$K = \frac{15 * 60}{\sum_{i=1}^{n} h_i}$$

Le modèle multiplie alors chacune des N valeurs h_i simulées par K. La somme des h_i ainsi corrigées vaudra exactement 15 minutes (temps de la simulation).