

RAPPORT

CETE de LYON

Centre d'Études
Techniques
de LYON

Département Mobilités

Affaire
6SRM01

Août 2013

Estimation du trafic sur les remontées mécaniques

Audit de la méthode actuelle et recherche d'améliorations

Août 2013

Ministère de l'Écologie, du Développement Durable et de l'Énergie



Estimation du trafic sur les remontées mécaniques

Audit de la méthode actuelle et recherche d'améliorations

Rapport

Août 2013

Date	Version	Commentaires
01/08/13	1	Rédaction initiale par Aurélie Bousquet et relecture spécialisée Nicolas Wagner, responsable unité SPAD
14/08/13	2	Relecture Olivier Richard
14/11/13	3	Corrections mineures suite à la découverte d'une incohérence dans les données source

CETE69_R2_DM_Rapport_Rev4



Département Mobilités
25, avenue François Mitterrand
CASE n°1
69674 BRON CEDEX
Tél.: +33 (0)4 72 14 31 24 - Fax.: +33 (0)4 72 14 31 20
Courriel : dmob.cete-lyon@developpement-durable.gouv.fr

Récapitulatif de l'affaire

Client : M. Jérôme CHAUVET
STRMTG
Domaine Universitaire
1461, Rue de la Piscine
38400 SAINT-MARTIN D'HERES

Objet de l'étude : Estimation du trafic sur les remontées mécaniques - Audit de la méthode actuelle et recherche d'améliorations

Résumé de la commande : Le STRMTG élabore chaque année un rapport d'accidentologie sur les remontées mécaniques qui nécessite la connaissance du nombre de personnes transportées sur l'ensemble du parc.
L'étude consiste à examiner la méthode actuellement utilisée pour estimer le trafic supporté par les remontées mécaniques et à proposer d'éventuelles améliorations de cette méthode en vue d'obtenir une méthode de calcul fiable, unique et reconnue par la profession (exploitants, DSF, STRMTG,...).

Référence dossier : Affaire 6SRM01

Offre : Proposition technique et financière envoyée en août 2012
Devis N° 21 2012 D 147 / 21 2013 D 151 envoyés en août 2012 et novembre 2013 respectivement

Accord client : 12/09/12

Diffusion/Archivage : Confidentiel – Documentation CETE de Lyon

Chargé d'affaire : Aurélie BOUSQUET –Département Mobilités –
Tél. +33 (0)4 72 14 31 24 / Fax +33 (0)4 72 14 31 20
Courriel : dmob.cete-lyon@developpement-durable.gouv.fr

Constitution de l'équipe : Aurélie Bousquet

Mots Clés : Estimation de trafic, Remontées Mécaniques, Régression statistique

ISRN :

Liste des destinataires

Contact	Adresse	Nombre - Type
M. Jérôme CHAUVET	STRMTG Domaine Universitaire 1461, Rue de la Piscine 38400 SAINT-MARTIN D'HERES	Envoi par courrier électronique
M. Christophe SION	STRMTG Domaine Universitaire 1461, Rue de la Piscine 38400 SAINT-MARTIN D'HERES	Envoi par courrier électronique

Conclusion – Résumé

Le STRMTG souhaite explorer les possibilités d'améliorer la fiabilité des estimations de trafic qu'il réalise chaque année et diffuse au sein de deux publications : une plaquette présentant les chiffres de trafics annuels et un « rapport accidents ». Les données de trafic sont remontées par les exploitants via Domaines Skiables de France, sur la base du volontariat des exploitants.

La méthode mise en œuvre jusqu'ici repose sur un redressement statistique basé sur un ratio du moment de puissance, quantité égale au produit de la dénivelée par la capacité horaire de l'installation. Ces coefficients sont appliqués par classe selon le massif et la catégorie de l'installation (88 classes au total). L'étude a montré que la proportionnalité entre les trafics et le moment de puissance, implicite lors de ce redressement, n'était pas vérifiée statistiquement. La principale variable explicative des trafics serait plutôt la capacité horaire de l'installation, qui présente une très forte corrélation avec la variable de trafic. De plus, la très forte segmentation des remontées en classes par massif et par catégories conduit à de faibles échantillons par classe (voire des échantillons nuls), dégradant la fiabilité des prévisions qui peuvent être réalisées au travers de cette méthode.

Un tour d'horizon des méthodes utilisées dans d'autres pays (Suisse, Autriche, Andorre et Italie) n'a pas permis de dégager de méthode alternative de traitement des non-répondants, ces pays ayant pour certains des données quasi-exhaustives (Andorre et Italie) ou mettant en œuvre des procédures de traitement des non-réponses très simplifiées (redressement global sur le nombre d'installations pour l'Autriche et traitement manuel des non-réponses avec les trafics des années précédentes ou des informations auxiliaires trouvées sur Internet pour la Suisse).

Une réflexion a été menée sur la segmentation des installations de remontées mécaniques en classes homogènes en termes de trafics supportés et d'attractivité des usagers. Une segmentation selon 6 classes distinctes a été proposée :

1. téléphériques,
2. télécabines,
3. télésièges débrayables ou fixes,
4. téléskis à perches débrayables ou fixes,
5. téléskis à câble bas,
6. autres installations (ascenseurs, funiculaires...).

En complément de la base FIRM fournie par le STRMTG et recensant les installations et leurs caractéristiques physiques, des fichiers de données complémentaires ont été fournis par DSF : trafic hivernal annuel, fréquentation en journées-skieurs vendues par station, nombre de jours d'ouverture de la remontée pendant la saison hivernale. Ces données complémentaires, remontées par les exploitants sur la base du volontariat, ne sont disponibles que sur un échantillon d'installations, cet échantillon pouvant être différent pour chaque variable.

Les méthodes mises en œuvre dans cette étude sont des régressions. Différents modèles de régression ont été testés : modèle linéaire, modèle de Poisson, modèle quasi-poissonien et modèle négatif binomial. Les modèles de Poisson et quasi-poissonien se sont révélés non adaptés aux données, puisque leur test de significativité statistique était toujours négatif.

Les modèles linéaires avec transformation de la variable à expliquer (le trafic) en racine carrée se sont révélés plus performants que les modèles négatifs binomiaux puisqu'ils présentent en moyenne des erreurs de prévision plus faibles (de 1 à 4 % en erreur relative), ces erreurs ayant été évaluées au travers d'un processus de validation croisée déployé sur les données de trafic disponibles sur 5 ans. C'est pourquoi, les modèles linéaires ont été finalement retenus pour effectuer les prévisions des valeurs de trafic manquantes.

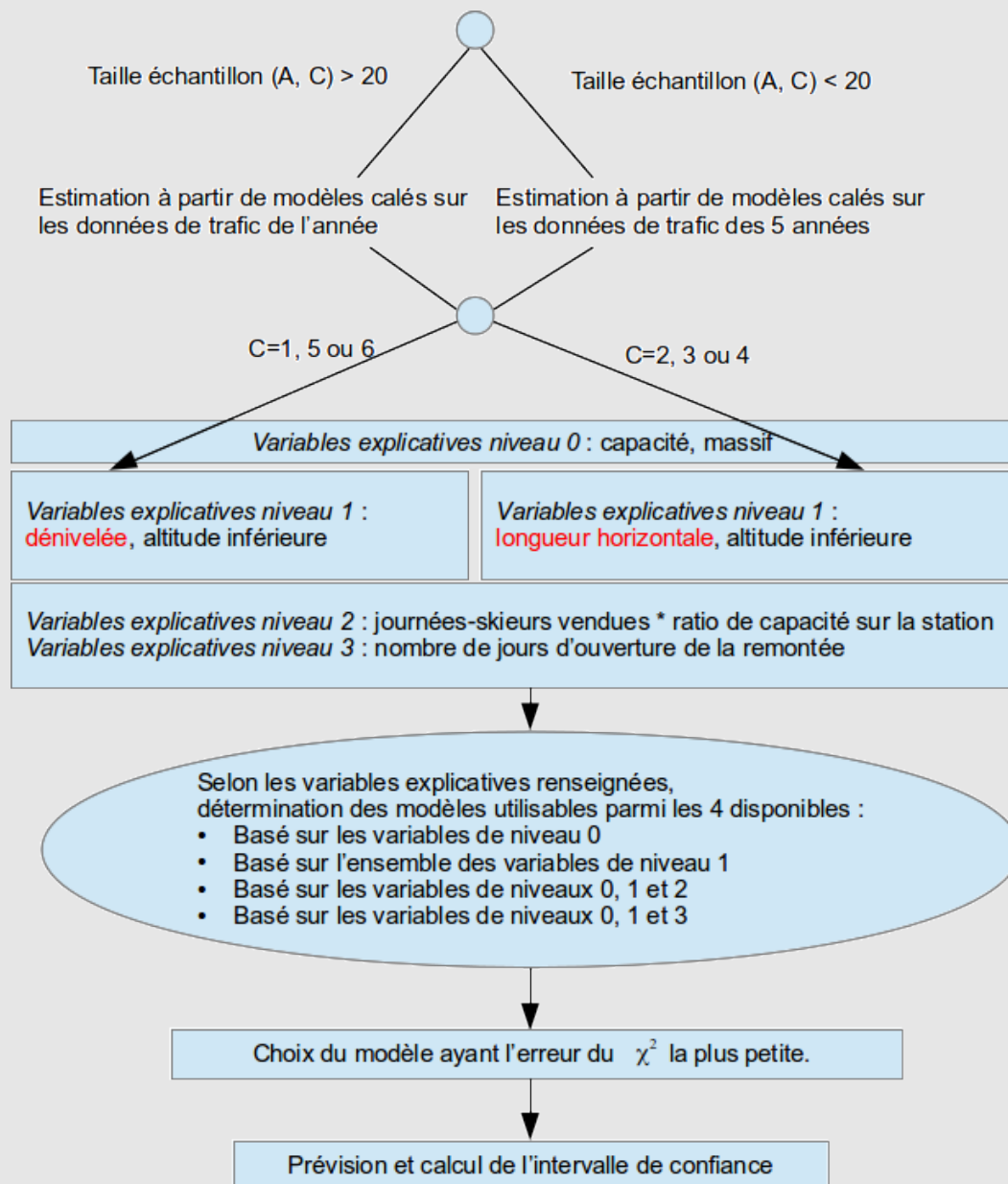
La base FIRM recensant les installations n'étant pas complète (valeurs manquantes sur certaines installations et certaines variables) et les données complémentaires de DSF n'étant disponibles que pour des échantillons d'installations, il a fallu adapter les modèles aux données. Pour cela 4 modèles distincts ont été proposés :

1. Modèle minimal : explique le trafic par l'appartenance à un massif et la capacité horaire de l'installation ;
2. Modèle basé sur les caractéristiques physiques des installations : explique le trafic par l'appartenance à un massif, la capacité horaire, la longueur horizontale ou la dénivelée (selon la catégorie d'installations) ainsi que l'altitude inférieure ;
3. Modèle basé sur les données de fréquentation : ajoute comme variable explicative au modèle 2 le produit entre le ratio de capacité de la remontée par rapport à sa station et le nombre de journées-skieurs vendues dans sa station ;
4. Modèle basé sur le nombre de journées d'ouverture de la remontée : ajoute comme variable explicative au modèle 2 le nombre de jours d'ouverture de la remontée.

On applique l'un ou l'autre selon les données disponibles pour la remontée et la précision respective des différents modèles, évaluée par validation croisée.

Dans le cas où l'échantillon de données de trafic pour la catégorie est trop faible pour l'année (<20), on utilise les 5 années de données pour caler le modèle. Cela « gomme » les effets d'évolution temporelle des trafics mais permet toutefois d'avoir une estimation qui correspond à un comportement moyen sur les dernières années.

La procédure d'estimation du trafic pour une remontée de catégorie C pour l'année A est résumée par le schéma ci-dessous.



Procédure de prévision de la valeur de trafic supporté par une installation appartenant à la catégorie C pendant l'année A

Une fois la prévision réalisée à l'échelle de la remontée, il faut ensuite agréger les prévisions et les trafics recensés pour obtenir un trafic sommé par catégorie d'installation ou par massif. Si la précision des estimations individuelles est assez faible en raison de la distribution statistique sous-jacente de la variable de trafic qui présente une très forte dispersion, la précision des estimations agrégées augmente sensiblement puisque les erreurs commises sur l'ensemble des précisions ont tendance à se compenser en moyenne. Toutefois, il faut se souvenir que plus la classe sur laquelle on agrège les données présente un faible effectif, plus le risque d'erreur est élevé.

Les résultats montrent que l'ancienne méthode d'estimation sous-estime systématiquement les trafics par rapport à la nouvelle pour les agrégations par massif, d'autant plus pour les petits massifs. Pour les agrégations par catégories, elle sous-estime également dans la majorité des cas les trafics. La nouvelle

méthode permet de fournir des estimations moyennes pour les massifs sur lesquels aucun trafic n'est renseigné, c'est le cas de la Corse. Les intervalles de confiance des trafics estimés avec la nouvelle méthode restent larges pour certaines catégories de remontées : les téléskis à câble bas, les téléphériques et les autres installations (funiculaires, ascenseurs). Ceci souligne l'intérêt qu'il y aurait à enrichir les données de trafic sur ces catégories. D'autre part, environ 20 remontées ne sont pas prises en compte dans l'estimation chaque année, faute de renseignements sur leurs caractéristiques dans la base FIRM. Renseigner complètement la base FIRM constitue également un angle d'amélioration de la fiabilité des estimations de trafic.

Les évolutions relatives d'une année sur l'autre sont du même ordre de grandeur avec l'ancienne et la nouvelle méthode.

Bron, le 14/08/2013

M. Olivier Richard
Responsable de l'unité 4

Sommaire

1 - DIAGNOSTIC DE LA MÉTHODE ACTUELLE.....	<u>9</u>
1.1 - Présentation de la méthode.....	<u>9</u>
1.2 - Analyse des effectifs et taux de sondage par classe d'installations.....	<u>9</u>
1.3 - Analyse de la nature du modèle.....	<u>13</u>
1.4 - Analyse de la qualité de la base de données.....	<u>14</u>
2 - AUTRES MÉTHODES UTILISÉES POUR L'ESTIMATION DES TRAFICS..	<u>15</u>
2.1 - Construction d'un indicateur de trafic sur les remontées mécaniques dans d'autres pays que la France.....	<u>15</u>
2.1.1 - Autriche.....	<u>15</u>
2.1.2 - Suisse.....	<u>15</u>
2.1.3 - Andorre et Italie.....	<u>16</u>
2.2 - Construction de l'indice de circulation sur le réseau routier national.....	<u>16</u>
3 - COMPARAISON DE PLUSIEURS MODÈLES DE RÉGRESSION POUR LA MISE AU POINT D'UNE NOUVELLE MÉTHODE D'ESTIMATION.....	<u>18</u>
3.1 - Analyse générale des données.....	<u>18</u>
3.1.1 - Segmentation des remontées en classes.....	<u>18</u>
3.1.2 - Examen des variables explicatives potentielles.....	<u>20</u>
3.1.3 - Analyse des corrélations.....	<u>21</u>
3.1.4 - Choix de la forme de la variable à expliquer pour le modèle linéaire.....	<u>23</u>
3.2 - Comparaison des modèles pour l'estimation des trafics.....	<u>24</u>
3.2.1 - Ajustement à partir des caractéristiques physiques des installations.....	<u>25</u>
3.2.2 - Ajustement à partir des caractéristiques physiques des installations et de données d'ouverture et de fréquentation des stations	<u>26</u>
3.2.3 - Analyse globale des résultats.....	<u>26</u>
4 - ESTIMATIONS ANNUELLES DES TRAFICS.....	<u>31</u>
4.1 - Procédure d'estimation.....	<u>31</u>
4.1.1 - Formules des modèles.....	<u>31</u>
4.1.2 - Analyse des résultats.....	<u>32</u>
4.2 - Calcul des agrégations de trafic.....	<u>33</u>
4.2.1 - Agrégations par massif.....	<u>33</u>
4.2.2 - Agrégations par catégories d'installations.....	<u>35</u>
4.2.3 - Trafic total annuel.....	<u>37</u>

5 - CONCLUSION.....	38
5.1 - Synthèse des résultats de l'étude.....	38
5.2 - Précaution d'utilisation des chiffres.....	39
5.3 - Pistes d'amélioration.....	39
RÉFÉRENCES.....	41
ANNEXE A – ÉLÉMENTS DE MÉTHODOLOGIE SUR LES RÉGRESSIONS..	42
Modèle linéaire.....	42
Modèle linéaire généralisé.....	44
Comparaisons de modèles.....	45
ANNEXE B – GRAPHIQUES DE CORRÉLATION DES VARIABLES EXPLICATIVES ET DE LA VARIABLE À EXPLIQUER.....	48
ANNEXE C – TRAITEMENTS RÉALISÉS SUR LES DONNÉES.....	49
ANNEXE D – CATÉGORIES D'INSTALLATIONS DE REMONTÉES MÉCANIQUES.....	50

1 - Diagnostic de la méthode actuelle

1.1 - Présentation de la méthode

La méthode actuelle mise en place par le STRMTG pour estimer les trafics annuels sur les remontées mécaniques consiste à réaliser un sondage sur un échantillon de remontées et à extrapoler un trafic global à partir des données de ce sondage. L'unité statistique est l'installation de remontée mécanique. L'extrapolation est réalisée par massif et par catégorie de remontée (22 catégories et 7 massifs, formant 88 classes d'installations au total).

Les données de trafic sont obtenues par interrogation des exploitants des remontées et fournies par ceux-ci sur la base du volontariat. Il est par conséquent impossible de définir un plan d'échantillonnage afin de contrôler la structure de l'échantillon final et il existe un risque de voir certaines catégories de remontées peu ou pas représentées.

Au-delà de mesures de relance de certains non-répondants qui pourraient améliorer la structure de l'échantillon, le principal levier pour améliorer la précision des estimations de trafic est donc de travailler sur la prévision des valeurs de trafic manquantes.

La méthode actuelle utilise un redressement par le ratio sur la base d'une variable de calage : le moment de puissance de l'installation. Cette variable auxiliaire est connue à l'échelle individuelle pour toutes les remontées mécaniques grâce à la base FIRM maintenue annuellement par le STRMTG. Elle est définie comme le produit de la dénivelée (capacité d'élévation de la remontée mécanique en mètres) par le débit maximum de l'installation (en nombre de passagers par heure) et se fournit donc en $m.passagers.h^{-1}$. Il s'agit d'une mesure d'efficacité de la remontée mécanique et on suppose que cette efficacité est directement liée à l'attractivité de la remontée pour les passagers. Le postulat implicite de la méthode actuellement utilisée est donc que la variable « moment de puissance » est corrélée positivement et linéairement avec le trafic annuel supporté par l'installation pour chaque classe issue du croisement entre la catégorie d'installation et le massif.

1.2 - Analyse des effectifs et taux de sondage par classe d'installations

Le redressement porte sur les 88 classes d'installations de la base FIRM. La taille des échantillons enquêtés conditionne la fiabilité de l'estimation de trafic pour une classe d'installations. On considère généralement que 30 valeurs est un minimum pour caler une relation statistique.

Nous présentons ici les tailles d'échantillon mises en regard avec les effectifs par classe afin de qualifier la qualité de l'estimation effectuée sur chacune de ces classes en 2012. La signification des codes de catégories d'installations est données en annexe D. On distingue trois situations :

- le taux de sondage (échantillon / effectif) de la classe est nul (cases rouges) : 32 classes ;

- l'échantillon est insuffisant, c'est-à-dire de taille inférieure à 30 et non exhaustif (cases oranges) : 42 classes ;
- l'échantillon est suffisant, c'est-à-dire qu'il est de taille au moins égale à 30 ou qu'il est exhaustif (cases vertes) : 14 classes.

Échantillon / effectif 2012	Alpes du Nord	Alpes du Sud	Corse	Jura	Massif Central	Pyrénées	Vosges	Total
TBA	2 / 3					0 / 1		2 / 4
TBP	0 / 2	0 / 2			0 / 1			0 / 5
TBV	14 / 30	3 / 4			2 / 3	1 / 12		20 / 49
DMD	4 / 8	1 / 1						5 / 9
DMV	1 / 4							1 / 4
TCD	61 / 100	7 / 12		2 / 2	0 / 1	4 / 14		74 / 129
TCP	8 / 9	1 / 2						9 / 11
TMV	4 / 5	2 / 3						6 / 8
TPM	0 / 1							0 / 1
TSD	125 / 230	18 / 39		2 / 3	2 / 2	15 / 27	1 / 3	163 / 304
TSCD	1 / 2	3 / 7		0 / 1				4 / 10
TSF	201 / 391	41 / 108		2 / 14	6 / 15	38 / 81	6 / 7	294 / 616
RDP	281 / 695	105 / 303	0 / 11	18 / 104	12 / 95	74 / 184	25 / 100	515 / 1492
RFP	49 / 150	9 / 57	0 / 2	1 / 11	1 / 9	1 / 32	1 / 18	62 / 279
RCAB	2 / 44	1 / 9		0 / 10	0 / 8	1 / 5		4 / 76
RCOB	3 / 194	1 / 60	0 / 1	1 / 31	0 / 11	0 / 32	0 / 15	5 / 344
RAC	1 / 35	0 / 10		0 / 4	0 / 1	0 / 9	0 / 2	1 / 61
ASC	2 / 5				0 / 1			2 / 6
FUN	3 / 8				0 / 2	0 / 2		3 / 12
EDS	1 / 1			0 / 1				1 / 2
TRSM (non pris en compte)	0 / 151	0 / 50		0 / 4	0 / 3	0 / 45	0 / 3	0 / 256
Total	811 / 2180	195 / 685	0 / 14	26 / 190	26 / 158	139 / 454	33 / 151	1230 / 3832

Part des effectifs par classe pour lesquels on a une valeur de trafic hivernal recensée en 2012

(vert : taille d'échantillon satisfaisante, orange : taille d'échantillon insuffisante, rouge : échantillon nul)

Les trafics des installations appartenant à des classes du premier type (cases rouges) seront totalement omises dans les estimations fournies. Sur les classes pour lesquelles le taux de sondage n'est pas nul, la méthode considère qu'il existe un modèle linéaire reliant le trafic et le moment de puissance des installations. Lorsque le nombre d'observations de la classe est trop petit (cases oranges), l'estimation réalisée pour cette classe est entachée d'une forte incertitude.

On sait cependant que les classes avec un faible échantillon sont généralement celles dont le moment de puissance total est faible. Bien que seules 16 % des classes aient des échantillons suffisants, elles représentent en fait 71 % des places offertes et 78 % du moment de puissance (variable sur laquelle est effectué le redressement). Les remontées complètement omises par le redressement (cases rouges) qui représentent 26 % des classes ne représentent que 3 % des places offertes et 1 % des moments de puissance.

	Nombre de classes	Places offertes	Moment de puissance (places.dénivelée)
Échantillon non nul (vert + orange)	56 / 88 = 64 %	3 643 219 / 3 765 065 = 97 %	959 371 / 968 908 = 99 %
Échantillon > 30 (vert)	14 / 88 = 16 %	2 676 275 / 3 765 065 = 71 %	760 392 / 968 908 = 78 %

Taux de sondage par classe, par places offertes et par places.dénivelée offertes pour 2012

% MTPU global	Alpes du Nord	Alpes du Sud	Autre	Corse	Jura	Massif Central	Pyrénées	Vosges	Total
ASC	0,02		0,00			0,00			0,03
DMD	1,44	0,06							1,50
DMV	0,13								0,13
EDS	0,00		0,01		0,00				0,01
FUN	0,90		0,06			0,01	0,03		1,00
RAC	0,04	0,01			0,00	0,00	0,02	0,00	0,08
RAE	1,06	0,13	0,00		0,04	0,05	0,09	0,01	1,38
RCAB	0,05	0,01	0,00		0,02	0,01	0,01		0,09
RCOB	0,06	0,02	0,00	0,00	0,01	0,00	0,01	0,01	0,11
RDP	8,51	4,46	0,00	0,10	0,93	0,83	2,26	1,00	18,10
RFP	0,23	0,09		0,00	0,01	0,01	0,04	0,03	0,41
TBA	0,25						0,00		0,25
TBP	0,03	0,08				0,02			0,12
TBV	1,59	0,14				0,10	0,14		1,98
TCD	10,00	0,87			0,21	0,15	1,39		12,62
TCP	0,09	0,03							0,12
TMV	0,02	0,03	0,00						0,05
TPM	0,01								0,01
TRSM	0,02	0,02			0,00	0,00	0,02	0,00	0,07
TSCD	0,24	0,69			0,06				0,99
TSD	25,20	4,12			0,31	0,13	2,71	0,17	32,64
TSF	18,66	4,54	0,02		0,46	0,55	3,82	0,26	28,31
Total	68,54	15,31	0,09	0,10	2,04	1,87	10,55	1,49	100,00

Pourcentage du moment de puissance total par classe croisé avec la taille de l'échantillon de données de trafic pour 2012 (vert : taille d'échantillon satisfaisante, orange : taille d'échantillon insuffisante, rouge : échantillon nul)

Le rapport « accidents » et la plaquette annuelle du STRMTG sur les trafics, diffusent des valeurs de trafic agrégées par massif et par catégorie d'installation. Si les pourcentages de moment de puissance représentés par les cases rouges et orange peuvent paraître limités à l'échelle globale des installations, lorsque l'on désagrège les données par massif et par catégorie, on s'aperçoit qu'ils sont nettement plus importants et introduisent une probabilité d'erreur élevée dans les estimations fournies.

Dans l'agrégation des trafics par massif, on relève notamment les problèmes suivants :

- pour les Alpes du Sud, 40 % environ du moment de puissance appartient à des classes sur lesquelles l'estimation n'est pas significative ;
- pour la Corse, le taux de sondage est nul, on ne peut donc pas fournir d'estimation ;
- pour le Massif Central, 10 % du moment de puissance est totalement omis, 83 % du moment de puissance appartient à des classes sur lesquelles l'estimation n'est pas significative.

% MTPU par massif	Alpes du Nord	Alpes du Sud	Corse	Jura	Massif Central	Pyrénées	Vosges
ASC	0,03				0,26		
DMD	2,10	0,41					
DMV	0,19						
EDS	0,00			0,02			
FUN	1,32				0,34	0,31	
RAC	0,06	0,08		0,17	0,05	0,15	0,15
RAE	1,55	0,84		1,90	2,89	0,90	0,43
RCAB	0,07	0,05		0,82	0,32	0,07	
RCOB	0,09	0,16	0,41	0,40	0,24	0,10	0,35
RDP	12,42	29,12	96,59	45,67	44,60	21,44	67,49
RFP	0,33	0,59	2,99	0,47	0,35	0,39	2,25
TBA	0,36					0,00	
TBP	0,04	0,53			0,86		
TBV	2,32	0,94			5,45	1,35	
TCD	14,58	5,71		10,36	8,08	13,16	
TCP	0,13	0,21					
TMV	0,04	0,17					
TPM	0,01						
TRSM	0,03	0,11		0,15	0,00	0,23	0,20
TSCD	0,34	4,53		2,73			
TSD	36,77	26,90		14,95	7,16	25,68	11,47
TSF	27,23	29,66		22,36	29,40	36,23	17,66
Total	100	100	100	100	100	100	100

Pourcentage du moment de puissance du massif par classe croisé avec les tailles d'échantillons de données de trafic pour 2012 (vert : taille d'échantillon satisfaisante, orange : taille d'échantillon insuffisante, rouge : échantillon nul)

Dans l'agrégation par catégorie d'installations, on relève notamment les points suivants :

- 32 % du moment de puissance des ascenseurs est totalement omis dans l'estimation ;
- pour les tapis, les taux de sondage sont nuls, on ne peut pas fournir d'estimation ;
- 73 % du moment de puissance des téléskis à câble bas appartient à des classes sur lesquelles l'estimation n'est pas significative, 27 % est totalement omis dans l'estimation.

% MTPU par catégorie d'équipement	Alpes du Nord	Alpes du Sud	Corse	Jura	Massif Central	Pyrénées	Vosges	Total
ASC	78,34	0,00	0,00	0,00	21,66	0,00	0,00	100
DMD	87,06	3,80	0,00	0,00	9,14	0,00	0,00	100
DMV	100,00	0,00	0,00	0,00	0,00	0,00	0,00	100
EDS	40,00	0,00	0,00	60,00	0,00	0,00	0,00	100
FUN	95,86	0,00	0,00	0,00	0,68	3,45	0,00	100
RAC	55,24	16,62	0,00	4,58	1,18	19,63	2,75	100
RAE	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100
RCAB	57,92	8,09	0,00	18,50	6,71	8,79	0,00	100
RCOB	54,54	20,65	0,33	7,16	3,91	8,91	4,50	100
RDP	47,41	24,57	0,54	5,09	4,56	12,31	5,52	100
RFP	55,24	22,25	0,71	2,28	1,62	9,93	7,97	100
TBA	100,00	0,00	0,00	0,00	0,00	0,00	0,00	100
TBP	20,53	66,38	0,00	0,00	13,09	0,00	0,00	100
TBV	81,77	7,43	0,00	0,00	5,26	5,54	0,00	100
TCD	80,31	7,26	0,00	1,65	0,00	10,78	0,00	100
TCP	73,55	26,45	0,00	0,00	0,00	0,00	0,00	100
TMV	48,36	51,64	0,00	0,00	0,00	0,00	0,00	100
TPM	100,00	0,00	0,00	0,00	0,00	0,00	0,00	100
TRSM	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100
TSCD	23,94	70,39	0,00	5,67	0,00	0,00	0,00	100
TSD	77,30	12,57	0,00	0,93	0,41	8,27	0,52	100
TSF	66,33	15,53	0,00	1,56	1,94	13,66	0,98	100

Pourcentage du moment de puissance de la catégorie par classe croisé avec les tailles d'échantillons de données de trafic pour 2012 (vert : taille d'échantillon satisfaisante, orange : taille d'échantillon insuffisante, rouge : échantillon nul)

Le fait de considérer un redressement différencié sur les 88 classes tend a priori à sous-estimer le trafic en raison des classes omises (cases rouges) et introduit une incertitude très forte sur l'estimation réalisée sur les classes à faible échantillon (cases oranges). Un regroupement des classes d'installations ayant a priori un comportement similaire en termes d'attractivité auprès des usagers paraît à ce stade être une piste d'amélioration de la fiabilité des estimations.

1.3 - Analyse de la nature du modèle

En complément des tests sur la taille des échantillons et donc sur la possibilité d'avoir des coefficients de redressement significatifs statistiquement, il faut aussi tester la validité de l'hypothèse de relation linéaire entre le trafic et le moment de puissance, qui est implicite dans la méthode actuelle. Pour cela, on se concentre sur les classes pour lesquelles l'échantillon est de taille supérieure ou égale à 30 puisque pour les autres, il n'est pas possible de tester la validité de la relation linéaire en raison du trop petit nombre d'observations.

	Alpes du Nord	Alpes du Sud	Pyrénées
RAE	R²=0.11 R ² ajusté=0.09 p-valeur test de Fisher de significativité du modèle=0.02		
RDP	R²=0.01 R ² ajusté=0.01 p-valeur test de Fisher de significativité du modèle=0.08	R²=0.03 R ² ajusté=0.02 p-valeur test de Fisher de significativité du modèle=0.08	R²=0.12 R ² ajusté=0.11 p-valeur test de Fisher de significativité du modèle=0.002
RF	R²=0.02 R ² ajusté=-0.01 p-valeur test de Fisher de significativité du modèle=0.4		
TCD	R²=0.23 R ² ajusté=0.21 p-valeur test de Fisher de significativité du modèle=0.0001		
TSD	R²=0.07 R ² ajusté=0.06 p-valeur test de Fisher de significativité du modèle=0.003		
TSF	R²=0.04 R ² ajusté=0.04 p-valeur test de Fisher de significativité du modèle=0.003	R²=0.02 R ² ajusté=-0.00 p-valeur test de Fisher de significativité du modèle=0.3	R²=0.19 R ² ajusté=0.17 p-valeur test de Fisher de significativité du modèle=0.006

Résultats des ajustements de relations linéaires entre le trafic hivernal et le moment de puissance de l'installation les classes ayant un échantillon d'au moins 30 valeurs en 2012

Parmi les modèles linéaires ayant été ajustés, 4 ne sont pas significatifs au seuil de 5 % (p-valeur du test de Fisher > 0.05). De plus, pour l'ensemble des modèles, le coefficient de détermination R² qui détermine la part de variance expliquée par le modèle est très faible (0,23 au maximum pour les télécabines débrayables dans les Alpes du Nord). Les graphiques ci-dessous illustrent la faible relation linéaire qui existe entre le moment de puissance et les trafics observés. La fiabilité des données extrapolées par le redressement sera donc faible.

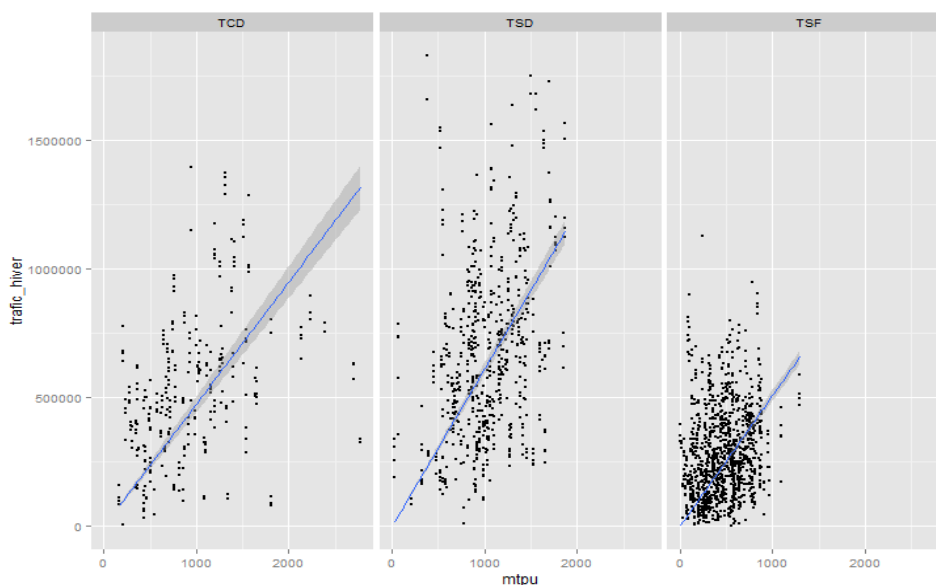


Illustration des ajustements linéaires et des données réelles pour les télécabines à attaches débrayables, les télésièges débrayables et les télésièges fixes des Alpes du Nord

1.4 - Analyse de la qualité de la base de données

Certaines données de trafic sont renseignées mais ne sont pas rattachées à une installation, généralement en raison d'une correspondance imparfaite entre le nom de l'installation et/ou de l'exploitant dans la base FIRM et dans le fichier des trafics obtenu via le questionnaire Internet envoyé aux exploitants. En 2012, 71 installations ont fait l'objet d'une déclaration et n'ont pas été incluses dans l'estimation.

À la suite de ce diagnostic et avant de proposer une méthode d'estimation des trafics alternative, nous allons examiner les méthodes statistiques utilisées pour réaliser des estimations de données de trafic, qu'il s'agisse des méthodes utilisées par les homologues du STRMTG à l'étranger pour estimer le trafic sur les remontées mécaniques ou de la méthode utilisée en France pour construire un indice national de circulation routière.

2 - Autres méthodes utilisées pour l'estimation des trafics

2.1 - Construction d'un indicateur de trafic sur les remontées mécaniques dans d'autres pays que la France

Dans le cadre de cette étude, les homologues étrangers du STRMTG ont été interrogés pour connaître leur méthode d'estimation des trafics sur les remontées mécaniques de leur pays. Une vingtaine de pays exploitant des remontées mécaniques fournissent annuellement des chiffres de trafic dans le cadre du congrès ITTAB (Internationale Tagung des Technischen Aufsichtsbehörden). Ce congrès réunit chaque année les services de contrôle des installations à câble transportant des personnes du monde entier.

Nous rapportons ici le contenu des retours de 4 pays européens qui ont été interrogés sur leur méthode d'estimation des trafics.

2.1.1 - Autriche

Le nombre d'installations est connu (y compris les ascenseurs) et mis à jour annuellement. Toutefois les informations détaillées sur les types d'installations ne sont pas disponibles.

Les exploitants des remontées mécaniques indiquent chaque année le nombre de courses réalisées : en montée/descente, été/hiver. Ces données ne sont pas toujours exhaustives mais le taux de réponse est bon. Les données manquantes sont extrapolées à partir d'un redressement basé sur le nombre total d'installations. Le trafic recensé est multiplié par le nombre total d'installations et divisé par le nombre d'installations pour lesquelles on a eu une réponse. Les agrégations par sous-ensemble géographique ne sont pas calculées.

2.1.2 - Suisse

L'Office Fédéral des Transports Suisse dispose de données de trafic sur toutes les installations par câble sous concession du pays (funiculaires, téléphériques bi-câbles, télécabines et télésièges débrayables). Les autres installations plus légères (petites installations de moins de 8 passagers et téléskis) sont contrôlées par les autorités cantonales.

La Suisse met en œuvre chaque année un sondage électronique concernant les transports publics, incluant les transports par câble gérés au niveau fédéral. Les exploitants font remonter leurs données de trafic issues de comptages électroniques ou estimées lorsqu'il n'existe pas de compteurs. Avec ce sondage, 95 % environ des installations voient leur trafic renseigné par les exploitants. Pour estimer les 5 % restants, des méthodes d'imputation de valeurs manquantes sont mises en œuvre : à partir des données de trafic des années précédentes, de rapports annuels d'activités ou d'indications publiées sur Internet par les opérateurs.

Des passagers.km sont ensuite calculés pour opérer des comparaisons avec d'autres systèmes de transport public.

2.1.3 - Andorre et Italie

En Andorre, le trafic est estimé pour toutes les remontées ayant un compteur automatique. En Italie, les 357 installations de remontées mécaniques possèdent toutes des compteurs. Par conséquent, les chiffres de trafic dont disposent les services de contrôle sont exhaustifs.

2.2 - Construction de l'indice de circulation sur le réseau routier national

Le problème d'estimation des trafics sur les remontées mécanique rappelle un problème voisin : l'estimation des trafics sur le réseau routier en vue de la construction d'un indice national de circulation.

L'indice de circulation du réseau routier national est calculé chaque année par le SETRA¹. C'est une estimation des distances parcourues sur le réseau routier national par l'ensemble des véhicules, mesurée en véhicules.km, qui est notamment utilisée afin d'alimenter les comptes annuels des transports.

La méthode de calcul est présentée de façon détaillée dans [1]. On calcule l'indice à partir d'un sondage réalisé sur un échantillon de stations de mesure automatique des trafics implantées sur le réseau routier non concédé. Le sondage sur le réseau concédé est exhaustif grâce aux données issues des péages.

L'indice est construit pour mesurer des évolutions de trafic depuis une année de référence. On dispose d'une première série de 1970 à 2005 qui concerne 35 000 km de réseau. Suite à la décentralisation d'une partie du réseau routier national, une nouvelle version de l'indice a été proposée en choisissant l'année 2001 comme référence. Cette nouvelle version de l'indice est basée sur 20 000 km de réseau.

La méthode consiste à établir une base d'indices en véhicules.km représentant les trafics moyens journaliers pour plusieurs sous-réseaux. On a défini pour cela un échantillon au moyen d'un plan de sondage stratifié par sous-réseau, par zone géographique (zones d'action des CETE) et par classe de taux de charge (rapport du trafic moyen journalier par la capacité théorique de la route).

Pour déterminer l'échantillon qui sera relevé chaque année, on réalise un recensement exhaustif des trafics sur le réseau concerné à l'année de référence (1970 pour la première vague et 2001 pour la seconde), on détermine les taux de charge et on les affecte à des classes. On détermine ensuite l'échantillon par stratification en posant comme contrainte que, pour l'année de référence, les véhicules.km reconstitués à partir des postes sondés soient strictement égaux aux véhicules.km recensés. Cet échantillon est fixe dans le temps. L'indice de circulation est donc estimé à réseau constant depuis l'année de référence.

1 Service d'études sur les transports, les routes et leurs aménagements

La taille de l'échantillon est définie de manière à avoir une précision comprise entre 2 et 4 % avec un intervalle de confiance à 95 %, autrement dit, il y a 95 % de chances que l'indice réel se situe dans un intervalle de +/- 2 à 4 % autour de l'indice calculé (variation selon les mois).

Pour pallier les éventuelles défaillances de certaines stations, l'échantillon strictement nécessaire pour atteindre cette précision a été augmenté d'une vingtaine de stations. On s'autorise ainsi à augmenter le coefficient de redressement associé à une zone, un sous-réseau et à une classe de taux de charge lorsqu'une station de cette zone, de ce sous-réseau et de cette classe de taux de charge était en panne pour l'année considérée.

L'échantillon de la seconde version de l'indice qui commence en 2001 comporte 350 stations (sur plus de 1 000 disponibles pour l'ensemble du réseau routier national). La méthode d'estimation est globalement inchangée par rapport à la première vague, seuls les sous-réseaux ont été redéfinis, avec le souci d'avoir une homogénéité des fonctions de la voirie et de la dynamique de croissance des trafics à l'intérieur de chaque sous-réseau [2].

La construction de cet indice de circulation, bien que présentant des similitudes avec notre problème à première vue, a en fait un objectif différent qui est de mesurer les évolutions au cours du temps des trafics supportés par les infrastructures, à réseau constant. Lorsqu'on l'utilise, on suppose donc implicitement que les évolutions de véhicules.km constatées sur les infrastructures appartenant au réseau choisi à l'année de référence sont similaires à celles constatées sur le reste du réseau mis en service depuis. On suppose également que la hiérarchie des voiries définie par leur taux de charge est conservée dans le temps. Ces simplifications sont faites en raison de la très grande complexité qui réside dans la mise en œuvre d'une estimation des trafics sur le réseau routier.

La méthode appliquée ici ne peut donc se transposer directement au cas des remontées mécaniques, pour lesquelles l'indicateur d'accidentologie nécessite de connaître non pas les trafics sur un réseau de référence constant dans le temps mais bien sur l'ensemble des installations, le parc et les caractéristiques des installations évoluant fortement au cours du temps et de façon non homogène sur l'ensemble du territoire.

Le diagnostic de la méthode actuellement utilisée nous a permis d'identifier un problème de fiabilité des prévisions, d'une part en raison de la forte segmentation en classes des installations et d'autre part parce que la relation de proportionnalité supposée entre le moment de puissance des installations et le trafic supporté n'est pas vérifiée sur les données dont on dispose. Dans le prochain paragraphe, nous allons examiner comment les méthodes de régression peuvent nous permettre d'estimer les valeurs manquantes de trafic et voir dans quelle mesure elles nous permettent d'augmenter la fiabilité des estimations réalisées.

3 - Comparaison de plusieurs modèles de régression pour la mise au point d'une nouvelle méthode d'estimation

Nous proposons une méthode basée sur les techniques de régression statistique qui font l'objet d'une abondante littérature (voir par exemple [3], [4], [5], [6], [7], [8], [9], [10], [11]). Les principaux éléments méthodologiques utilisés dans le cadre de cette étude font l'objet d'une présentation en annexe A.

3.1 - Analyse générale des données

3.1.1 - Segmentation des remontées en classes

Comme on l'a vu précédemment, une trop forte segmentation des installations conduit à des résultats non significatifs sur les classes pour lesquelles on a trop peu de données de trafic recensées pour caler un modèle. Ainsi, il est préférable de réduire le nombre de classes.

Les échanges avec le STRMTG ont amené à retenir 8 classes qui correspondent à une segmentation ayant du sens pour l'analyse de l'accidentologie :

- téléphériques bicâbles,
- télécabines,
- télésièges débrayables,
- télésièges fixes,
- téléskis à perche débrayable,
- téléskis à attache fixe,
- téléskis à câble bas et
- autres installations.

Nous avons choisi, dans le cadre de l'estimation des trafics, de **regrouper les télésièges débrayables et les télésièges fixes ainsi que les téléskis à perche débrayable et les téléskis à attache fixe**. En effet, le caractère débrayable d'une installation change essentiellement sa vitesse d'exploitation et on tient déjà compte de cette notion au travers de l'indicateur de capacité horaire, puisqu'une installation débrayable aura une capacité horaire plus importante qu'une installation équivalente non-débrayable. Ces regroupements sont opérés pour le **calage des modèles d'estimation des trafics**. Rien n'empêchera d'estimer par la suite séparément les trafics sur les installations débrayables et non débrayables (mais avec le même modèle), afin de disposer de chiffres d'accidentologie pondérés par les trafics sur chacune de ces catégories.

- Catégorie 1 : *téléphériques bicâbles*
Regroupement de TBA, TBP et TBV
- Catégorie 2 : *télécabines*
Regroupement de DMD, DMV, TCD, TCP, TMV et TPM
- Catégorie 3 : *télésièges débrayables et fixes*
Regroupement de TSCD, TSD et TSF
- Catégorie 4 : *téléskis à perche débrayable et à attache fixe*

- Regroupement de RDP, RAE et RFP
- Catégorie 5 : *téléskis à câble bas*
Regroupement de RAC, RCOB et RCAB
- Catégorie 6 : *autres installations*
Regroupement de ASC, EDS et FUN

Parmi les installations qui figurent dans la base FIRM, certaines ont été exclues des estimations de trafic.

- *Les tapis* : On constate (sur 2012 comme sur les 4 années précédentes) qu'il n'y a aucune valeur de trafic relevée pour ces installations. Un travail complémentaire pour élaborer un modèle des trafics supportés par les tapis pourra être réalisé ultérieurement dans le cas où des données de trafic recensé deviendraient disponibles.
- *Les installations classés en massif « Autres »* : Il s'agit dans la très grande majorité d'installations de transport par câble urbaines pour lesquelles les estimations de trafic sortent largement du cadre de cette étude qui porte sur les remontées mécaniques de stations de montagne.
- *Les installations identifiées comme « appareil de service »* : Elles ne supportent a priori qu'un trafic très faible et ne se comportent donc pas comme les autres installations.
- *Les installations dont l'exploitation est suspendue pour la saison.*

Remarque : On trouve dans la base FIRM actuelle des remontées dont la date de mise en service est antérieure à l'année d'estimation des trafics et qui ne figurent pourtant pas dans les fichiers d'inventaire des installations ayant été utilisés par le STRMTG pour estimer les trafics en 2008, 2009, 2010, 2011 et 2012. Le STRMTG précise qu'il s'agit de remontées dont la déclaration a été faite a posteriori par les exploitants et qui ont donc été intégrées dans la base FIRM après leur mise en service effective. Par conséquent, elles n'ont pas été prises en compte dans l'estimation des trafics mise en œuvre jusqu'ici par le STRMTG et nous conserverons pour le test de la nouvelle méthode la liste de remontées utilisée par l'ancienne méthode afin de rendre comparables les résultats obtenus.

Les échantillons 2012 et les effectifs de la base FIRM par classe sont donnés par le tableau suivant.

Échantillon / Effectif	Alpes du Nord	Alpes du Sud	Corse	Jura	Massif Central	Pyrénées	Vosges	Total
1 Téléphériques	16 / 35	3 / 6			2 / 4	1 / 13		22 / 58
2 Télécabines	78 / 127	11 / 18		2 / 2	0 / 1	4 / 14		95 / 162
3 Télésièges	327 / 623	72 / 154		4 / 18	8 / 17	53 / 108	7 / 10	461 / 930
4 Téléskis à attaches fixes ou débrayables	378 / 957	117 / 378	0 / 13	19 / 120	16 / 110	80 / 226	26 / 121	636 / 1925
5 Téléskis à câble bas	6 / 224	2 / 129	0 / 1	1 / 49	0 / 23	1 / 91	0 / 20	10 / 737
6 Autres installations	6 / 14			0 / 1	0 / 3	0 / 2		6 / 20
Total	811 / 2180	195 / 685	0 / 14	26 / 190	26 / 158	139 / 454	33 / 151	1230 / 3832

Volume de l'échantillon de trafics et effectif total des différentes classes d'installations pour 2012

Avec ces nouvelles catégories, les effectifs des classes sont globalement plus satisfaisants même s'il reste des classes pour lesquelles aucune ou un nombre très faible d'observations de trafic sont recensées. Les marges contiennent toutes des observations, sauf celles correspondant au massif « Corse ».

3.1.2 - Examen des variables explicatives potentielles

Il s'agit de choisir parmi les variables disponibles lesquelles sont significatives pour expliquer la valeur de trafic supportée par une installation. Les variables pourront être considérées sous leur forme simple ou transformée (passage au log, racine carrée...). On dispose de données de trois types :

- **les variables quantitatives qui sont connues pour toutes les remontées mécaniques, tous les ans (issues de la base FIRM)² :**
 - Capacité horaire : en passagers/heure
 - Altitude inférieure de la remontée : en mètres
 - Dénivelée : en mètres
 - Longueur horizontale : en mètres
 - Moment de puissance : en mètres.passagers/heure
- **les variables qualitatives qui sont connues pour toutes les remontées mécaniques, tous les ans (issues de la base FIRM) :**
 - Massif
 - Catégorie d'installation (les 6 catégories définies ci-dessus)
- **les variables quantitatives qui sont connues seulement pour un échantillon de remontées mécaniques, l'échantillon étant modifié chaque année et potentiellement différent de l'échantillon sur lequel sont connus les trafics (données remontées par les exploitants via DSF) :**
 - Nombre de journées-skieurs vendues par station
 - Nombre de jours d'ouverture de chaque installation.

A partir de ces variables, nous avons calculé des variables supplémentaires qui nous paraissent a priori pouvoir expliquer le trafic.

L'indicateur « journées-skieurs vendues » donné à la station peut être *réparti* entre les différentes remontées de la station. Pour cela, les variables *ratio_capa* / *ratio_mtpu* / *ratio_deni* ont été introduites. Elles représentent respectivement le ratio entre la capacité / le moment de puissance / la dénivelée de la remontée et la somme des capacités / moments de puissance / dénivelées de toutes les remontées de la station à laquelle appartient la remontée. Trois variables supplémentaires ont ensuite été introduites qui correspondent au produit du nombre de journées-skieurs vendues par chacun de ces trois ratios.

Plutôt que de considérer l'altitude inférieure de la remontée comme variable expliquant l'attractivité de celle-ci pour les skieurs, nous nous sommes demandé si la différence d'altitude entre le bas de la remontée et la station ne pourrait pas être une meilleure variable explicative. Ces deux variables sont toutefois fortement corrélées linéairement et nous devons donc choisir l'une ou l'autre dans les modèles.

2 La base FIRM actuelle présente des installations pour lesquelles certaines des variables citées ci-dessus ne sont pas connues. Les modèles qui s'appuient sur des variables explicatives avec valeurs manquantes excluront les installations concernées de l'estimation de trafic finale. Il y a donc un enjeu particulièrement fort à renseigner les variables citées ci-dessus dans la base FIRM.

3.1.3 - Analyse des corrélations

On réalise une analyse de corrélation linéaire et de corrélation au sens des rangs de Spearman (mesure de la corrélation globale, pas seulement de la corrélation linéaire) des variables explicatives potentielles avec le trafic hivernal ainsi qu'entre elles, sur les 5 ans de données disponibles. Cette analyse donne une première indication sur l'importance des différentes variables explicatives et permet d'identifier les différences de comportement par rapport aux variables explicatives par catégorie d'installations.

Il en ressort tout d'abord que la capacité horaire est la variable la plus fortement corrélée au trafic hivernal, pour toutes les catégories d'installations.

Variables explicatives		Capacité horaire	Dénivelée	Moment de puissance	Altitude inférieure	Différence d'altitude avec la station	Longueur horizontale	Journées skieurs vendues * ratio dénivelée	Journées skieurs vendues * ratio capacité	Journées skieurs vendues * ratio mtpu	Jours d'ouverture hivernale de la remontée
Téléphériques	Nombre d'observations	120	120	120	120	120	89	54	54	54	114
	Corrélation linéaire	0.77	0.14	0.63	0.22	-0.02	-0.17	0.02	0.16	0.04	0.31
	Corrélation de rang (Spearman)	0.75	0.21	0.44	0.35	-0.05	0.03	0.17	0.5	0.33	0.28
Télécabines	Nombre d'observations	534	534	534	534	534	360	242	242	242	502
	Corrélation linéaire	0.66	0.24	0.55	0.34	-0.04	0.42	0.08	0.05	0.11	0.41
	Corrélation de rang (Spearman)	0.7	0.25	0.57	0.4	0	0.4	0.3	0.44	0.42	0.46
Télesièges	Nombre d'observations	2693	2693	2693	2686	2686	2048	1333	1333	1333	2524
	Corrélation linéaire	0.75	0.15	0.59	0.2	0.03	0.31	0.11	0.25	0.17	0.57
	Corrélation de rang (Spearman)	0.73	0.15	0.56	0.19	0.01	0.31	0.24	0.39	0.36	0.65
Téléskis à attache fixe ou débrayables	Nombre d'observations	4149	4149	4149	4064	4064	2616	1998	1998	1998	3967
	Corrélation linéaire	0.13	-0.11	-0.08	0.17	0.03	-0.04	-0.02	0.12	-0.02	0.67
	Corrélation de rang (Spearman)	0.14	-0.09	-0.06	0.19	0.03	-0.02	0.16	0.28	0.17	0.72
Téléskis à câble bas	Nombre d'observations	72	71	71	63	63	42	39	39	39	68
	Corrélation linéaire	0.69	-0.14	0.07	0.36	0.01	0.17	0.05	0.18	0.06	0.5
	Corrélation de rang (Spearman)	0.43	-0.03	0.29	0.38	0.03	0.14	0.31	0.46	0.41	0.6
Autres	Nombre d'observations	32	32	32	32	32	19	19	19	19	32
	Corrélation linéaire	0.83	0.58	0.8	-0.38	0.18	0.56	0.58	0.49	0.59	0.65
	Corrélation de rang (Spearman)	0.82	0.52	0.62	-0.37	0.09	0.56	0.4	0.67	0.43	0.65

Mesures de corrélation entre la variable à expliquer et les variables explicatives sur 5 années (2008 à 2012)

Dans les régressions, il est nécessaire d'éviter les colinéarités entre variables explicatives pour limiter les instabilités du modèle (voir annexe A). Pour chaque couple de variables fortement corrélées linéairement, on doit choisir de garder, pour chaque catégorie d'installations, uniquement celle qui est la plus corrélée avec le trafic hivernal. Ici, les variables fortement corrélées linéairement sont (voir en annexe B les graphiques de croisement par paires des variables) :

- la dénivelée et la longueur horizontale,
- l'altitude inférieure et la différence d'altitude avec la station,
- le moment de puissance et la capacité,
- le moment de puissance et la dénivelée,
- la longueur horizontale et le moment de puissance,
- les trois indicateurs basés sur le nombre de journées-skiEUR vendues.

Le moment de puissance semble une bonne variable explicative du trafic, toutefois, il est fortement corrélé linéairement à la capacité et à la dénivelée (par définition). La capacité est plus fortement corrélée que le moment de puissance au trafic, donc **on préférera conserver la capacité que le moment de puissance dans les modèles.**

La différence d'altitude entre le bas de la remontée et la station est peu corrélée avec les trafics hivernaux, en tout cas toujours moins que l'altitude inférieure de la remontée. **On retient donc l'altitude inférieure comme variable explicative.**

Selon les catégories, c'est la dénivelée ou la longueur horizontale qui est la plus corrélée avec le trafic. Les *téléskis à attaches fixes ou débrayables* sont **plus sensibles à la dénivelée**. Les *télécabines, télésièges et téléskis à câble bas* sont **plus sensibles à la longueur horizontale**. Les *téléphériques* et *autres installations* sont **sensibles aux deux variables**, nous testerons donc laquelle amène au meilleur ajustement. On remarque l'impact attractif de la dénivelée pour les *téléphériques* et les *autres installations* et son impact répulsif (signe négatif du coefficient de corrélation) pour les *téléskis*.

Enfin, parmi les trois indicateurs basés sur les journées skiEUR vendues par station, on constate que **c'est celui basé sur le ratio de capacité de la remontée qui est le plus fortement corrélé au trafic recensé**. C'est donc celui qu'on choisit de conserver.

Enfin, très logiquement, **le nombre de jours d'ouverture hivernal est fortement corrélé au trafic recensé**, en particulier pour les *téléskis* et les *autres installations*. Lorsqu'elle est disponible, cette variable devrait donc pouvoir être avantageusement utilisée pour compléter le modèle et améliorer sa précision.

On constate que les *téléskis fixes et débrayables* se comportent très différemment des autres types d'installations, notamment de par leur **réaction négative à la dénivelée** et leur **faible réaction à la capacité horaire** en comparaison des autres types d'installations.

3.1.4 - Choix de la forme de la variable à expliquer pour le modèle linéaire

Les résultats du test de normalité de Shapiro-Wilk (voir annexe A) appliqué à la variable de trafic hivernal pour les 5 années de recensement confondues sont présentés ci-dessous.

P-valeur du test de Shapiro-Wilk	Y	Log Y	$\sqrt{Y}$
Téléphériques	2.086e-10	2.63e-10	0.02072
Télécabines	< 2.2e-16	3.183e-15	0.0001092
Télesièges	< 2.2e-16	< 2.2e-16	1.789e-11
Téléskis fixe et débrayables	< 2.2e-16	< 2.2e-16	4.934e-08
Téléskis à câble bas	9.562e-08	7.317e-06	0.01301
Autres	0.002973	0.0007545	0.00355

Les variables Y ainsi que log(Y) conduisent dans la très grande majorité des cas à rejeter l'hypothèse nulle selon laquelle la distribution est gaussienne (*p-valeur* très faible). La variable $\sqrt{Y}$ donne dans la très grande majorité des cas une *p-valeur* supérieure à celles obtenues pour Y et log(Y), qui témoigne du fait que pour ces catégories d'installations c'est $\sqrt{Y}$ qui s'approche le plus d'une loi normale. C'est par conséquent la formulation que l'on retiendra pour la variable à expliquer.

Remarque : L'hypothèse d'indépendance des observations de trafic des modèles de régression n'est pas vérifiée : en effet, les trafics sur plusieurs remontées géographiquement proches sont liés (même nombre de forfaits vendus, même météo...). Cela signifie qu'en réalité la quantité d'information contenue dans les données est moindre que ce que l'on suppose (certaines données étant corrélées, le nombre de degrés de liberté diminue), ce qui amène en conséquence à sous-estimer la taille des intervalles de confiance calculés par le modèle.

3.2 - Comparaison des modèles pour l'estimation des trafics

Comme on l'a vu précédemment, les données issues de comptages suivent fréquemment une distribution de Poisson. C'est pourquoi nous avons testé les modèles suivants qui s'adaptent a priori bien à ce type de distribution :

- **modèle linéaire avec une transformation en racine carrée** dont on a vu au paragraphe précédent qu'elle normalisait mieux les données de trafics que le logarithme ;
- **modèle quasi-poissonien** (modèle linéaire généralisé) avec fonction de lien logarithmique ;
- **modèle négatif binomial** avec fonction de lien logarithmique.

Afin de limiter le nombre de modèles à ajuster et de permettre un calage sur des échantillons de taille plus importante, nous avons choisi de définir seulement 6 modèles, un par catégorie d'installations traitant l'ensemble des massifs. La distinction des massifs se fait à l'intérieur de chaque modèle en intégrant comme variable explicative le massif. Concrètement, la variable qualitative « massif » se traduit par 7 variables quantitatives indicatrices (une variable par massif valant 0 ou 1 pour chaque installation). On cherche donc à calibrer un modèle moyen pour tous les massifs et 7 valeurs additionnelles de trafic, chacune correspondant à un massif.

Ainsi, on attribuera à une remontée appartenant à un massif d'échantillon de taille nulle (par exemple le massif Central pour la catégorie 2 en 2012) le comportement « moyen » des remontées de sa catégorie sur l'ensemble des massifs. On commet forcément une erreur en faisant une telle approximation, toutefois cela nous permet de fournir une estimation pour un massif sur lequel aucune valeur de trafic n'a été recensée pour l'année courante.

Une remontée appartenant à un massif ayant un échantillon de taille non nulle verra son trafic ajusté d'une constante par rapport au comportement moyen sur l'ensemble des massifs. Plus l'échantillon du massif sera petit (par exemple le Jura pour la catégorie 2 en 2012), plus l'erreur commise sera forte, toutefois l'erreur finale sera moins importante que si on avait essayé de caler un modèle spécifique pour le massif (il serait calé ici sur 2 valeurs seulement !).

Comme précisé précédemment, un bon calage des modèles de régression nécessite d'avoir 10 à 20 observations minimum par variable explicative incluse dans le modèle, soit un nombre d'observations global de *(nombre de variables explicatives + nombre d'indicatrices de massif)*10* ainsi que 10 observations minimum par massif afin que le calage du coefficient de chaque indicatrice de massif soit bon.

Pour la base FIRM de 2012, le tableau ci-dessous résume les échantillons qui auraient été souhaitables pour un modèle à 4 variables explicatives (le maximum utilisé dans le cadre de l'étude) + les indicatrices de massif.

	Nombre d'indicateurs de massif pour la catégorie	Nombre d'observations minimum par massif dans la base	Échantillon minimum souhaité tous massifs confondus	Nombre d'observations tous massifs confondus dans la base
1 Téléphériques	4	1	80 (seulement 58 installations de ce type)	22
2 Télécabines	5	0	90	95
3 Télésièges	6	4	100	461
4 Téléskis à attaches fixes ou débrayables	7	0	110	636
5 Téléskis à câble bas	7	0	110	10
6 Autres	4	0	80 (seulement 20 installations de ce type)	6

Comparaison des échantillons minimum souhaités et des échantillons présents dans la base 2012

Cette première analyse des échantillons annuels souligne qu'il y aurait un intérêt à relancer les exploitants des installations de catégories 1, 2, 5 et 6 pour obtenir des données de trafic plus nombreuses. Les catégories 1 et 6 semblent particulièrement problématiques puisqu'un taux de sondage de 100 % ne permettrait pas encore de caler un modèle correctement. Sur ces catégories, il paraît donc nécessaire à ce stade de viser un recueil de données de trafic exhaustif de manière à ne pas avoir à estimer les trafics par un modèle.

3.2.1 - Ajustement à partir des caractéristiques physiques des installations

Un premier ajustement des modèles a été réalisé uniquement à partir des variables explicatives contenues dans la base FIRM, relatives aux caractéristiques physiques des installations. On se restreint aux variables qui ont été identifiées comme pertinentes par l'analyse des corrélations présentée ci-dessus (massif, capacité horaire, altitude inférieure et dénivelée ou longueur horizontale).

Ces variables n'étant pas toutes renseignées dans certains enregistrements de la base, un premier modèle utilisant un minimum de variables explicatives a été testé. Il est basé sur la capacité horaire et le massif. Un modèle plus complet a ensuite été ajusté à l'aide d'une méthode de sélection de variables pas à pas. Les résultats de ces ajustements sont résumés dans le tableau ci-après³.

Quelle que soit la catégorie d'installations considérée, les modèles quasi-poissonien se sont révélés non significatifs, ils ont donc été abandonnés. En revanche, le modèle négatif binomial est significatif dans la majorité des cas et le modèle linéaire l'est tout le temps. Les mesures d'erreurs RMSE, RMSE.Log et χ^2 (voir annexe A pour leur définition) sont globalement meilleures pour le modèle linéaire que pour le modèle négatif binomial. Comme on pouvait s'y attendre, l'ajout de variables supplémentaires permet d'améliorer les qualités prédictives du modèle, puisque cela fait diminuer les indicateurs d'erreurs.

3 Lorsque le modèle est non significatif, il apparaît sur fond orange, sinon il apparaît sur fond vert. Un seuil de significativité $\alpha=0,05$ a été retenu.

Remarque sur les *téléphériques* : le modèle minimal est identique au modèle complet. Seuls la capacité et le massif interviennent pour expliquer le trafic supporté par l'installation. Un test a été réalisé en introduisant la dénivelée puis un autre en la remplaçant par la longueur horizontale. Aucune des deux variables n'a été retenue par la procédure de sélection pas à pas.

Remarque sur les *téléphériques*, les *téléskis à câble bas* et les *autres installations* : Malgré leurs très faibles échantillons, il ne paraît pas raisonnable de les regrouper avec une autre catégorie étant donné leur comportement spécifique. Pour les *téléphériques* et les *autres installations*, nous avons vu qu'il était nécessaire de viser l'exhaustivité dans le recueil, le nombre d'installations existantes ne pouvant permettre de caler correctement un modèle de régression. Pour les *téléskis à câble bas*, il est nécessaire de fournir un effort particulier pour améliorer de façon sensible les échantillons.

3.2.2 - Ajustement à partir des caractéristiques physiques des installations et de données d'ouverture et de fréquentation des stations

Nous introduisons maintenant des variables explicatives supplémentaires au sein de deux modèles distincts :

- le nombre de journées-skieurs vendues multiplié par le ratio de capacité de la remontée par rapport à l'ensemble des remontées de la station : cette variable est ajoutée dans le modèle sous sa forme logarithmique, puisque parmi les transformations classiques, c'est celle qui nous permet de nous approcher le plus d'une loi normale ;
- le nombre de jours d'ouverture de la remontée : les tests réalisés ont amené à considérer une relation quadratique entre le trafic et le nombre de jours d'ouverture de la remontée, on ajoute donc au modèle les deux variables $jours.hiver$ et $jours.hiver^2$.

Les résultats pour ces deux familles de modèles sont également regroupés dans le tableau ci-dessous. Ici non plus le modèle quasi-poissonien n'est jamais significatif et ne figure donc pas dans le tableau.

L'introduction des variables de jours d'ouverture et de fréquentation des stations fait clairement diminuer l'erreur quadratique moyenne : on estime donc un peu mieux la dispersion autour de la moyenne.

3.2.3 - Analyse globale des résultats

L'analyse des sorties graphiques (résidus, valeurs observées/valeurs prévues) et des indicateurs d'erreurs consignés dans le tableau ci-dessous nous donne des éclairages sur la fiabilité des ajustements utilisés.

Une prévision de trafic fournie par le modèle correspond à la moyenne des trafics que l'on peut attendre pour une installation avec des caractéristiques données. On constate en pratique que, comme cela est souvent le cas pour des données de comptage, la dispersion autour de cette moyenne est très forte : elle est même supérieure à la moyenne des observations pour un niveau donné. L'estimation au

moyen d'une régression est donc très délicate si on se situe au niveau d'une remontée isolée mais devient plus envisageable lorsque le but est d'estimer une somme de trafic globale pour une classe de remontées. L'erreur du χ^2 le montre bien : pour une valeur isolée, elle est très élevée et du même ordre de grandeur que l'erreur pour la somme. Autrement dit, on se trompe autant au sens du χ^2 en évaluant la somme des trafics sur une classe qu'en évaluant le trafic d'une remontée isolée.

Cela est confirmé par les très faibles erreurs relatives sur la somme des installations de chaque catégorie : inférieure à 1 % pour les catégories dont les échantillons sont de bonne taille, plus élevée (jusqu'à 4 %) pour les catégories 5 et 6 à échantillons très faibles.

On en déduit qu'il est nécessaire de travailler sur des classes d'installations relativement agrégées, les effectifs dans chaque classe devant être importants afin que le modèle soit bien spécifié et que le fait de sommer « gomme » les erreurs d'estimation individuelles dues à la dispersion naturelle des données autour de la moyenne.

À l'exception des catégories 1, 5 et 6, pour lesquelles les échantillons sont relativement petits, le modèle linéaire apparaît toujours meilleur en termes de qualité de prévision que le modèle négatif binomial. Pour les catégories 1, 5 et 6, les résultats sont nuancés, donnant la préférence à l'un ou l'autre modèle selon les situations. Afin de simplifier les traitements et puisqu'il est raisonnable de penser que le modèle linéaire aurait également l'avantage si les échantillons étaient plus importants, **on choisit d'utiliser les prévisions réalisées grâce au modèle linéaire pour la mise en œuvre de la méthode**. Toutefois, le modèle linéaire ne permettant pas de calculer des intervalles de confiance pour la somme, en raison de la transformation en racine carrée appliquée à la variable, on cherchera à comparer les intervalles de confiance générés par le modèle négatif binomial et les prévisions générées par le modèle linéaire.

Attention : Les coefficients de détermination R^2 des modèles linéaires ne doivent pas être interprétés directement comme des mesures de qualité des modèles. En effet, les forts R^2 constatés pour les catégories 5 et 6 s'expliquent essentiellement par un phénomène mécanique lié aux faibles échantillons dans ces catégories. Le faible R^2 de la catégorie 4 (téléskis) s'explique en partie aussi par le très gros échantillon de données, dans lequel la variance à expliquer est forcément plus importante. Les coefficients de détermination ne doivent donc être analysés que pour comparer différents modèles à l'intérieur d'une même catégorie, pour un échantillon donné, de la même manière que les indicateurs AIC et BIC.

	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial
Méthode choix variables	Modèle minimal basé sur la capacité		Sélection de variables pas à pas Caractéristiques physiques des installations		Sélection de variables pas à pas Caractéristiques physiques des installations et données de fréquentation		Sélection de variables pas à pas Caractéristiques physiques des installations et nombre de jours d'ouverture	
Variable expliquée	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log
1. Téléphériques 48 observations								
Variables explicatives	sqrt(capacite)	massif offset(log(capacite))	sqrt(capacite)	massif offset(log(capacite))	sqrt(capacite)	massif offset(log(capacite))	sqrt(alt_inf) sqrt(capacite) jours_hiver l(jours_hiver^2)	massif offset(log(capacite))
Mesures d'adéquation du modèle	P-valeur Fisher = 3.50e-09 R² ajusté = 0,53	Test d'adéquation : 0.28	P-valeur Fisher = 3.50e-09 R² ajusté = 0,53	Test d'adéquation : 0.28	P-valeur Fisher = 3.50e-09 R² ajusté = 0,53	Test d'adéquation : 0.28	P-valeur Fisher = 1.74e-10 R² ajusté = 0,66	Test d'adéquation : 0.28
RMSE	168887	153694	168887,00	153694,00	168887	153694	139517	153694
RMSELog	0,45	0,38	0,45	0,38	0,45	0,38	0,45	0,38
χ^2	84279.9378	69529.3023	84279.9378	69529.3023	84279.9378	69529.3023	72907.824	69529.3023
Erreur Relative Moyenne sur la somme	-0.016	0.046	-0.016	0.046	-0.016	0.046	-0.003	0.046
$\Sigma \chi^2$	83234.533	81133.218	83234.533	81133.218	83234.533	81133.218	87209.62	81133.218
2. Télécabines 149 observations								
Variables explicatives	sqrt(capacite) massif	massif offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	massif offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	sqrt(alt_inf) log(is_ratio_capa) massif offset(log(capacite))	sqrt(alt_inf) sqrt(capacite) jours_hiver massif	sqrt(alt_inf) sqrt(lhor) jours_hiver l(jours_hiver^2) massif offset(log(capacite))
Mesures d'adéquation du modèle	P-valeur Fisher = 9.82e-23 R² ajusté = 0,51	Test d'adéquation : 0.22	P-valeur Fisher = 9.43e-26 R² ajusté = 0,57	Test d'adéquation : 0.22	P-valeur Fisher = 9.44e-26 R² ajusté = 0,57	Test d'adéquation : 0.21	P-valeur Fisher = 1.89e-30 R² ajusté = 0,62	Test d'adéquation : 0.17
RMSE	169924	192690	156945	192690	156945	189665	152954	206744
RMSELog	0.55	0.58	0.51	0.58	0.51	0.59	0.42	0.48
χ^2	92931.2312	110553.8313	94709.2688	110553.8313	94709.2688	106704.4502	75191.3196	106156.9617
Erreur Relative Moyenne sur la somme	-0.003	-0.069	0.000	-0.069	0.000	-0.060	0.004	-0.071
$\Sigma \chi^2$	71014.88	98396.791	79991.967	98396.791	79991.967	99120.642	91138.142	157481.63

	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial
Méthode choix variables	Modèle minimal basé sur la capacité		Sélection de variables pas à pas Caractéristiques physiques des installations		Sélection de variables pas à pas Caractéristiques physiques des installations et données de fréquentation		Sélection de variables pas à pas Caractéristiques physiques des installations et nombre de jours d'ouverture	
Variable expliquée	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log
3. Télésièges 932 observations								
Variables explicatives	sqrt(capacite) massif	massif offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	sqrt(alt_inf) sqrt(lhor) massif offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	sqrt(alt_inf) sqrt(lhor) log(js_ratio_capa) massif offset(log(capacite))	sqrt(lhor) sqrt(capacite) l(jours_hiver^2) massif	sqrt(lhor) jours_hiver l(jours_hiver^2) massif offset(log(capacite))
Mesures d'adéquation du modèle	P-valeur Fisher = 8.65e-175 R² ajusté = 0,58	Test d'adéquation : 0.08	P-valeur Fisher = 6.24e-186 R² ajusté = 0,61	Test d'adéquation : 0.08	P-valeur Fisher = 6.24e-186 R² ajusté = 0,61	Test d'adéquation :0.07	P-valeur Fisher = 3.53e-235 R² ajusté = 0,69	Test d'adéquation : 0.15
RMSE	178949	196759	172612	181410	172612	177106	159311	155895
RMSELog	0.64	0.68	0.61	0.64	0.61	0.63	0.46	0.43
χ^2	81541.2953	94699.9782	78150.3161	83246.2405	78150.3161	80372.2789	59267.1497	57554.499
Erreur Relative Moyenne sur la somme	-0.002	0.059	-0.002	0.038	-0.002	0.028	-0.002	-0.003
$\Sigma \chi^2$	85226.28	151925.109	84635.36	122742.801	84635.36	93647.13	78154.821	69715.399
4. Téléskis à attache fixe ou débrayable 1842 observations								
Variables explicatives	sqrt(capacite) massif	massif offset(log(capacite))	sqrt(alt_inf) sqrt(denivelee) sqrt(capacite) massif	sqrt(alt_inf) sqrt(denivelee) massif offset(log(capacite))	sqrt(alt_inf) sqrt(denivelee) log(js_ratio_capa) sqrt(capacite) massif	sqrt(alt_inf) sqrt(denivelee) log(js_ratio_capa) massif offset(log(capacite))	sqrt(capacite) jours_hiver massif	sqrt(denivelee) jours_hiver l(jours_hiver^2) massif offset(log(capacite))
Mesures d'adéquation du modèle	P-valeur Fisher = 7.56e-42 R² ajusté = 0,10	Test d'adéquation : 0.002	P-valeur Fisher = 4.01e-75 R² ajusté = 0,18	Test d'adéquation : 0.004	P-valeur Fisher = 7.60e-86 R² ajusté = 0,20	Test d'adéquation : 0.00	P-valeur Fisher = 0.000 R² ajusté = 0,56	Test d'adéquation : 0.05
RMSE	75907	77171	72563	75669	71685	75416	56897	59972
RMSELog	1.01	1.03	0.95	0.96	0.93	0.95	0.49	0.49
χ^2	51037.067	53139.1687	45567.2071	47594.865	43861.7441	46681.1916	25384.1127	27991.1265
Erreur Relative Moyenne sur la somme	0.001	-0.028	0.001	-0.026	0.001	-0.032	0.000	-0.031
$\Sigma \chi^2$	41342.222	49219.87	41537.414	54222.21	40030.563	54221.609	24354.111	36736.412

	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial	Modèle Linéaire	Modèle Négatif Binomial
Méthode choix variables	Modèle minimal basé sur la capacité		Sélection de variables pas à pas Caractéristiques physiques des installations		Sélection de variables pas à pas Caractéristiques physiques des installations et données de fréquentation		Sélection de variables pas à pas Caractéristiques physiques des installations et nombre de jours d'ouverture	
Variable expliquée	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log	$\sqrt{\text{trafic.hiver}}$	trafic_hiver, lien log
5. Téléskis à câble bas 22 observations								
Variables explicatives	sqrt(capacite)	offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	sqrt(alt_inf) sqrt(lhor) massif offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	sqrt(alt_inf) sqrt(lhor) log(js_ratio_capa) massif offset(log(capacite))	sqrt(alt_inf) sqrt(lhor) sqrt(capacite) massif	sqrt(alt_inf) sqrt(lhor) jours_hiver l(jours_hiver^2) massifAlpes.du.Nor d offset(log(capacite))
Mesures d'adéquation du modèle	P-valeur Fisher = 1.50e-05 R² ajusté = 0,60	Test d'adéquation : 0.27119	P-valeur Fisher = 2.59e-08 R² ajusté = 0,88	Test d'adéquation : 0.21	P-valeur Fisher = 2.59e-08 R² ajusté = 0,88	Test d'adéquation : 0.17	P-valeur Fisher = 2.59e-08 R² ajusté = 0,88	Test d'adéquation : 0.13
RMSE	104448	121086	54907	103783	54907	123471	54907	70062
RMSELog	1.28	1.32	0.28	0.21	0.28	0.26	0.28	0.19
χ^2	86359.4614	81810.8782	23133.8326	46307.5856	23133.8326	55888.9091	23133.8326	31691.5786
Erreur Relative Moyenne sur la somme	-0.011	0.078	-0.029	-0.05	-0.029	-0.013	-0.029	-0.027
$\Sigma \chi^2$	40799.224	55641.973	22282.021	74019.441	22282.021	89733.716	22282.021	35419.464
6. Autres installations 19 observations								
Variables explicatives	sqrt(capacite)	offset(log(capacite))	sqrt(alt_inf) sqrt(capacite)	sqrt(alt_inf) offset(log(capacite))	sqrt(alt_inf) sqrt(capacite)	sqrt(alt_inf) offset(log(capacite))	sqrt(alt_inf) sqrt(capacite)	sqrt(alt_inf) offset(log(capacite))
Mesures d'adéquation du modèle	P-valeur Fisher = 1.48e-06 R² ajusté = 0,74	Test d'adéquation : 0.35	P-valeur Fisher = 2.28e-07 R² ajusté = 0,83	Test d'adéquation : 0.30	P-valeur Fisher = 2.28e-07 R² ajusté = 0,83	Test d'adéquation : 0.30	P-valeur Fisher = 2.28e-07 R² ajusté = 0,83	Test d'adéquation : 0.30
RMSE	146474	128929	104138	108485	104138	108485	104138	108485
RMSELog	0.25	0.27	0.22	0.16	0.22	0.16	0.22	0.16
χ^2	57251.777	51871.0532	36919.5114	38002.2782	36919.5114	38002.2782	36919.5114	38002.2782
Erreur Relative Moyenne sur la somme	0.04	0.121	0.014	0.013	0.014	0.013	0.014	0.013
$\Sigma \chi^2$	65184.408	49764.725	44174.519	47896.861	44174.519	47896.861	44174.519	47896.861

Ajustement des modèles de régression du trafic hivernal, erreurs calculées par validation croisée

4 - Estimations annuelles des trafics

Dans cette partie, nous proposons une nouvelle méthode d'estimation des trafics mise en œuvre sous R, logiciel de statistiques libre et gratuit. La procédure de mise en œuvre est présentée dans une notice spécifique, jointe au présent rapport.

4.1 - Procédure d'estimation

Ayant déterminé les modèles à choisir à partir des 5 années de données, nous allons maintenant ajuster des modèles annuels. En effet, il est souhaitable de pouvoir se baser sur une estimation annuelle du trafic pour analyser les accidents. Pour cela, on utilise une procédure dérivée de celle mise en œuvre au paragraphe précédent :

- Seuls les modèles linéaires sont utilisés puisqu'on a vu qu'ils donnaient de meilleurs résultats que les autres en prévision.
- Les données utilisées en entrée sont en priorité les données de l'année. Ceci permet de tenir compte de paramètres conjoncturels qui agissent sur la demande mais qu'on ne mesure pas : météo et contexte économique⁴. Lorsque les échantillons de l'année sont trop faibles pour la catégorie, on utilise la base historique sur 5 ans pour déterminer une estimation de trafic. Ce procédé gomme les effets d'évolution temporels mais permet toutefois d'avoir une prévision correspondant à la moyenne des valeurs observées les années précédentes.
- Pour chaque modèle, on utilise la base d'apprentissage la plus grande possible. Ainsi la hiérarchie des 4 modèles (modèle minimal, modèle utilisant les caractéristiques physiques, modèle utilisant les données de fréquentation et modèle utilisant les jours d'ouverture) ne sera plus forcément respectée : le modèle minimal pourra donner des prévisions plus précises car il sera calé sur un plus grand nombre d'observations que les modèles plus élaborés qui utilisent plus de variables et ont donc des échantillons d'apprentissage plus petits. Pour chaque remontée, on choisit, parmi les modèles qui reposent sur des variables explicatives disponibles pour cette remontée, celui ayant l'erreur du χ^2 sur la somme la plus petite, celle-ci étant évaluée par une validation croisée. Le diagramme présenté en page 5 résume ce processus.

4.1.1 - Formules des modèles

Les 4 modèles utilisés sont basés sur les formules suivantes :

1. Modèle minimal, basé sur la capacité

$$\sqrt{\text{trafic.hiver}} = \beta_0 + \beta_1 \sqrt{\text{capacite}} + \beta_2 1_{\text{massifAlpesNord}} + \beta_3 1_{\text{massifAlpesSud}} + \dots$$

2. Modèle basé sur les caractéristiques physiques de la remontée

$$\sqrt{\text{trafic.hiver}} = \beta_0 + \beta_1 \sqrt{\text{capacite}} + \beta_2 \sqrt{\text{alt.inf}} + \beta_3 \sqrt{\text{denivelée}} + \beta_4 1_{\text{massifAlpesNord}} + \beta_5 1_{\text{massifAlpesSud}} + \dots$$

OU

$$\sqrt{\text{trafic.hiver}} = \beta_0 + \beta_1 \sqrt{\text{capacite}} + \beta_2 \sqrt{\text{alt.inf}} + \beta_3 \sqrt{\text{longueur.horizontale}} + \beta_4 1_{\text{massifAlpesNord}} + \beta_5 1_{\text{massifAlpesSud}} + \dots$$

⁴ Le nombre de journées-skieurs vendues nous informe sur le contexte économique, toutefois, il n'est disponible que pour certaines stations.

3. Modèle basé sur les données de fréquentation par station

$js.ratio.capa = \text{journées} - \text{skieurs vendues} * \text{ratio de capacité}$

$$\sqrt{\text{trafic.hiver}} = \beta_0 + \beta_1 \sqrt{\text{capacite}} + \beta_2 \sqrt{\text{alt.inf}} + \beta_3 \sqrt{\text{denivelée}} + \beta_4 \log(js.ratio.capa) + \beta_5 1_{\text{massifAlpesNord}} + \beta_6 1_{\text{massifAlpesSud}} + \dots$$

4. Modèle basé sur le nombre de jours d'ouverture de la remontée

$$\sqrt{\text{trafic.hiver}} = \beta_0 + \beta_1 \sqrt{\text{capacite}} + \beta_2 \sqrt{\text{alt.inf}} + \beta_3 \sqrt{\text{denivelée}} + \beta_4 \text{jours.ouv} + \beta_5 \text{jours.ouv}^2 + \beta_6 1_{\text{massifAlpesNord}} + \beta_7 1_{\text{massifAlpesSud}} + \dots$$

4.1.2 - Analyse des résultats

L'analyse par année a confirmé les meilleures performances des modèles linéaires par rapport aux négatifs binomiaux. Les tailles variables d'échantillon selon le modèle utilisé amènent comme prévu à privilégier parfois un modèle ayant moins de variables qu'un autre, son erreur du χ^2 sur la somme étant plus faible en raison du plus grand nombre d'observations utilisées pour le calage.

On observe que pour les *téléskis à câble bas* et les *autres installations* les échantillons annuels sont trop faibles pour permettre une estimation. Par conséquent, les échantillons sur 5 ans sont utilisés pour le calage. C'est également le cas des *téléphériques* en 2011. Lorsqu'il est possible de faire autrement, caler les modèles sur les données de 5 ans est bien entendu à éviter, d'une part parce que les évolutions annuelles sont gommées, car on effectue une moyenne des comportements des installations sur la période et d'autre part parce que l'échantillon ainsi formé reprend souvent plusieurs fois la même installation (avec par conséquent les mêmes valeurs de variables explicatives portant sur les caractéristiques physiques) pour des années différentes. Ainsi l'échantillon de calage n'est pas très varié, les plages de valeurs de calage pour les différentes variables explicatives sont limitées. **Pour éviter ces imprécisions, il serait nécessaire de renforcer les échantillons des catégories 1, 5 et 6, voire de tendre vers un recueil exhaustif.**

Catégorie	Hiver 2007-2008	Hiver 2008-2009	Hiver 2009-2010	Hiver 2010-2011	Hiver 2011-2012
1 Téléphériques	Annuel = 20	Annuel = 31	Annuel = 29	Annuel = 18 Sur 5 ans = 120	Annuel = 22
2 Télécabines	Annuel = 94	Annuel = 129	Annuel = 122	Annuel = 94	Annuel = 95
3 Télésièges	Annuel = 570	Annuel = 675	Annuel = 590	Annuel = 403	Annuel = 455
4 Téléskis à attaches fixes ou débrayables	Annuel = 940	Annuel = 1043	Annuel = 930	Annuel = 600	Annuel = 636
5 Téléskis à câble bas	Annuel = 14 Sur 5 ans = 72	Annuel = 19 Sur 5 ans = 72	Annuel = 17 Sur 5 ans = 72	Annuel = 12 Sur 5 ans = 72	Annuel = 10 Sur 5 ans = 72
6 Autres	Annuel = 7 Sur 5 ans = 32	Annuel = 6 Sur 5 ans = 32	Annuel = 7 Sur 5 ans = 32	Annuel = 6 Sur 5 ans = 32	Annuel = 6 Sur 5 ans = 32

Taille de l'échantillon du modèle minimal pour la base annuelle et pour la base sur 5 ans quand la base annuelle n'est pas suffisante

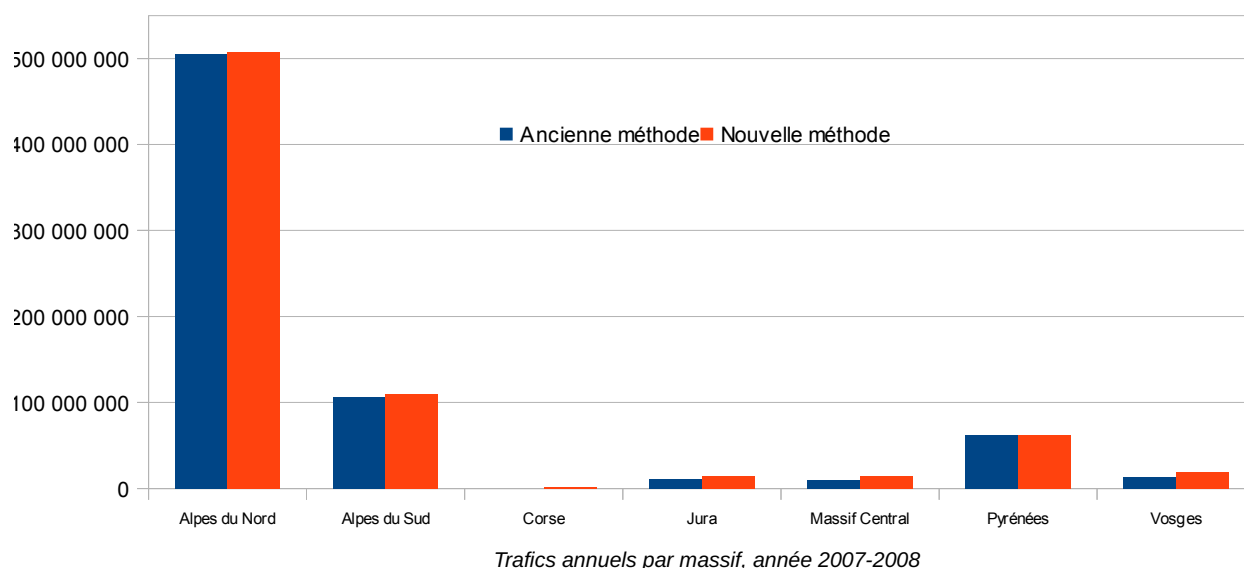
4.2 - Calcul des agrégations de trafic

Des chiffres de trafic sont publiés par le STRMTG par type d'installation ou par massif. Aucun chiffre de trafic croisant ces deux variables n'est affiché dans le rapport annuel publié sur le site du STRMTG ni dans le rapport « accident ». Nous ne descendons donc pas à ce niveau de finesse pour les estimations fournies par la nouvelle méthode, ce qui permet de plus de limiter les problèmes de fiabilité des résultats.

4.2.1 - Agrégations par massif

Le tableau ci-dessous reprend les estimations calculées grâce à l'ancienne et la nouvelle méthode, par massif. Les estimations par la nouvelle méthode sont accompagnées d'un intervalle de confiance approximatif⁵. On constate, comme prévu, que l'ancienne méthode avait tendance à plutôt sous-estimer le trafic par massif, d'autant plus fortement pour les massifs les moins fréquentés comme le Jura et le Massif Central. De plus, le trafic de Corse était systématiquement annulé faute de données observées sur ce massif, tandis que par la nouvelle méthode il est estimé selon le comportement moyen des installations similaires dans d'autres massifs. Par conséquent, le trafic estimé pourrait ne pas être très représentatif du massif Corse, toutefois c'est le mieux que l'on puisse faire en l'absence de données de trafic recensées sur ce massif.

Les installations non prises en compte correspondent à celles pour lesquelles il serait nécessaire de compléter la base FIRM pour disposer de toutes les variables explicatives nécessaires (capacité, altitude inférieure, etc.). Il y en a une vingtaine chaque année.



5 Le modèle linéaire avec transformation de la variable à expliquer par une racine carrée ne permet pas de calculer des intervalles de confiance pour les sommes. Par conséquent, les intervalles présentés ici ont été calculés à l'aide des variances estimés par le modèle négatif binomial équivalent. Il s'agit donc d'une approximation de l'intervalle de confiance réel. De plus, on sait que la non-indépendance des observations de trafic induit une sous-estimation systématique de la taille des intervalles de confiance.

	Massif	Alpes du Nord	Alpes du Sud	Corse	Jura	Massif Central	Pyrénées	Vosges
2008	Trafic mesuré	278 514 062	61 942 476	0	5 219 407	3 533 643	23 537 702	5 762 050
	Trafic estimé, nouvelle méthode	228 699 429	48 365 001	971 483	9 488 701	11 316 856	39 063 691	12 775 790
	Trafic total, nouvelle méthode	507 213 490	110 307 478	971 483	14 708 108	14 850 499	62 601 393	18 537 840
	Intervalle de confiance approximé à 95%	[505 318 917 ; 509 108 064]	[109 860 223 ; 110 754 732]	[878 321 ; 1 064 644]	[14 350 993 ; 15 065 223]	[14 569 349 ; 15 131 648]	[61 913 288 ; 63 289 498]	[17 042 006 ; 20 033 675]
	Taille de l'intervalle de confiance	+/-1 %	+/-1 %	+/-10 %	+/-4 %	+/-2 %	+/-2 %	+/-12 %
	Échantillon annuel	1 020	335	0	55	45	149	41
	Installations estimées	1 137	352	10	125	114	279	128
2009	Installations non estimées ⁶	0	0	0	20	0	0	3
	Trafic total, ancienne méthode	505 535 266	106 679 397	0	11 173 798	9 330 043	61 628 044	13 108 241
	Trafic mesuré	366 653 796	62 084 846	0	6 539 247	5 612 488	28 379 702	9 947 761
	Trafic estimé, nouvelle méthode	133903633	46927953	1 039 550	12 498 387	16 700 113	42 700 518	12 149 855
	Trafic total, nouvelle méthode	500 557 429	109 012 799	1 039 550	19 037 634	22 312 601	71 080 220	22 097 616
	Intervalle de confiance approximé à 95%	[499 715 122 ; 501 399 735]	[108 598 198 ; 109 427 399]	[1 016 063 ; 1 063 036]	[18 911 968 ; 19 163 299]	[21 994 870 ; 22 630 332]	[70 524 899 ; 71 635 540]	[20 279 704 ; 23 915 528]
	Taille de l'intervalle de confiance	+/-1 %	+/-1 %	+/-2 %	+/-1 %	+/-2 %	+/-1 %	+/-15 %
2010	Échantillon annuel	1 281	323	0	51	41	148	60
	Installations estimées	780	326	10	113	116	271	89
	Installations non estimées	0	0	0	12	0	0	2
	Trafic total, ancienne méthode	493 447 877	105 295 818	0	12 593 478	14 697 101	68 606 377	20 134 987
	Trafic mesuré	299 265 777	43 855 071	0	4 968 537	1 593 294	26 298 013	7 241 226
	Trafic estimé, nouvelle méthode	162837942	57 698 249	485 118	9 261 498	15 475 394	30 509 864	12 245 720
	Trafic total, nouvelle méthode	462 103 719	101 553 320	485 118	14 230 035	17 068 688	56 807 877	19 486 946
2011	Intervalle de confiance approximé à 95%	[460 202 980 ; 464 004 458]	[101 045 801 ; 102 060 838]	[469 010 ; 501 226]	[14 129 754 ; 14 330 316]	[16 821 001 ; 17 316 376]	[56 194 036 ; 57 421 718]	[17 432 239 ; 21 541 653]
	Taille de l'intervalle de confiance	+/-1 %	+/-1 %	+/-3 %	+/-1 %	+/-2 %	+/-2 %	+/-17 %
	Échantillon annuel	1 155	244	0	50	21	181	44
	Installations estimées	875	390	6	110	134	236	100
	Installations non estimées	1	0	0	12	0	0	2
	Trafic total, ancienne méthode	456 503 272	100 260 703	0	9 703 033	8 804 639	53 420 672	13 954 530
	Trafic mesuré	214 496 160	32 405 164	0	1 689 719	4 602 496	8 012 908	6 067 267
2012	Trafic estimé, nouvelle méthode	231 289 915	80 591 009	647 573	11 446 507	16 272 429	61 657 056	17 535 647
	Trafic total, nouvelle méthode	445 786 075	112 996 173	647 573	13 136 226	20 874 925	69 669 964	23 602 914
	Intervalle de confiance approximé à 95%	[443 881 299 ; 447 690 852]	[112 426 537 ; 113 565 808]	[625 456 ; 669 691]	[12 896 702 ; 13 375 750]	[20 679 438 ; 21 070 411]	[69 002 356 ; 70 337 572]	[20 816 173 ; 26 389 656]
	Taille de l'intervalle de confiance	+/-1 %	+/-1 %	+/-3 %	+/-2 %	+/-1 %	+/-1 %	+/-16 %
	Échantillon annuel	814	179	0	20	35	54	31
	Installations estimées	1 226	457	6	149	120	356	123
	Installations non estimées	1	0	0	18	0	0	2
2013	Trafic total, ancienne méthode	460 599 734	114 591 698	0	8 714 609	12 430 874	47 805 494	15 230 077
	Trafic mesuré	233 325 905	31 161 097	0	2 689 771	3 148 423	26 133 002	6 233 476
	Trafic estimé, nouvelle méthode	243 277 962	66 231 149	1 301 634	12 773 907	12 660 466	43 604 554	15 866 583
	Trafic total, nouvelle méthode	476 603 867	97 392 246	1 301 634	15 463 678	15 808 889	69 737 556	22 100 059
	Intervalle de confiance approximé à 95%	[474 698 839 ; 478 508 894]	[96 776 284 ; 98 008 209]	[1 265 464 ; 1 337 803]	[15 346 076 ; 15 581 279]	[15 532 339 ; 16 085 440]	[69 074 188 ; 70 400 924]	[19 597 202 ; 24 602 916]
	Taille de l'intervalle de confiance	+/-1 %	+/-1 %	+/-3 %	+/-1 %	+/-2 %	+/-2 %	+/-16 %
	Échantillon annuel	807	193	0	26	26	139	33
2014	Installations estimées	1 193	434	11	132	119	258	104
	Installations non estimées	2	0	0	12	0	0	2
	Trafic total, ancienne méthode	487 562 315	95 106 199	0	9 714 235	10 569 192	60 605 712	13 899 431

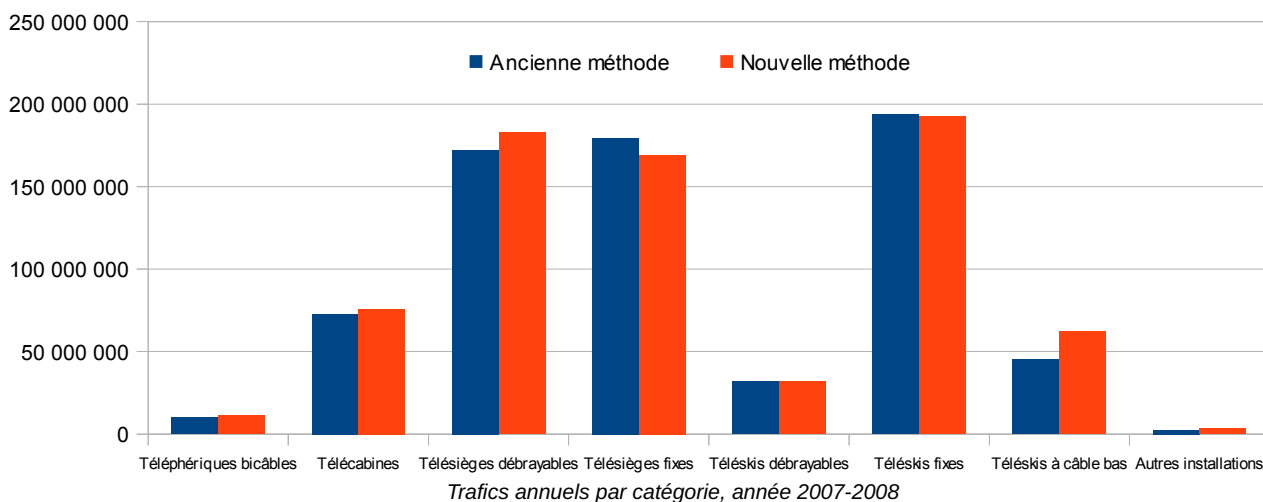
On remarque que le trafic sur le massif des Vosges est particulièrement mal estimé avec un intervalle de confiance représentant environ 15 % en plus ou en moins par rapport à la valeur estimée.

Le tableau ci-dessus permet d'identifier les pistes d'amélioration suivantes :

- L'amélioration de la base FIRM en renseignant plus complètement les variables explicatives permettrait que toutes les installations sont prises en compte dans l'estimation et que le meilleur modèle en termes de précision puisse être utilisé.
- Les trafics sur le massif Vosgien semblent présenter de fortes variabilités conduisant à de larges intervalles de confiance. Il y aurait donc un enjeu à augmenter le nombre de données de trafic relevées chaque année sur ce massif.
- Étant donné le petit nombre d'installations sur le massif Corse, un recueil exhaustif des données de trafic annuel semble réalisable et permettrait de fournir un chiffre vraiment représentatif du massif, le chiffre actuellement fourni correspondant à une moyenne de comportement des autres massifs sur des installations similaires.

4.2.2 - Agrégations par catégories d'installations

Le tableau ci-dessous présente les résultats agrégés par catégorie d'installations. Comme pour les massifs, bien que moins systématique, on remarque une tendance à la sous-estimation des trafics par l'ancienne méthode.

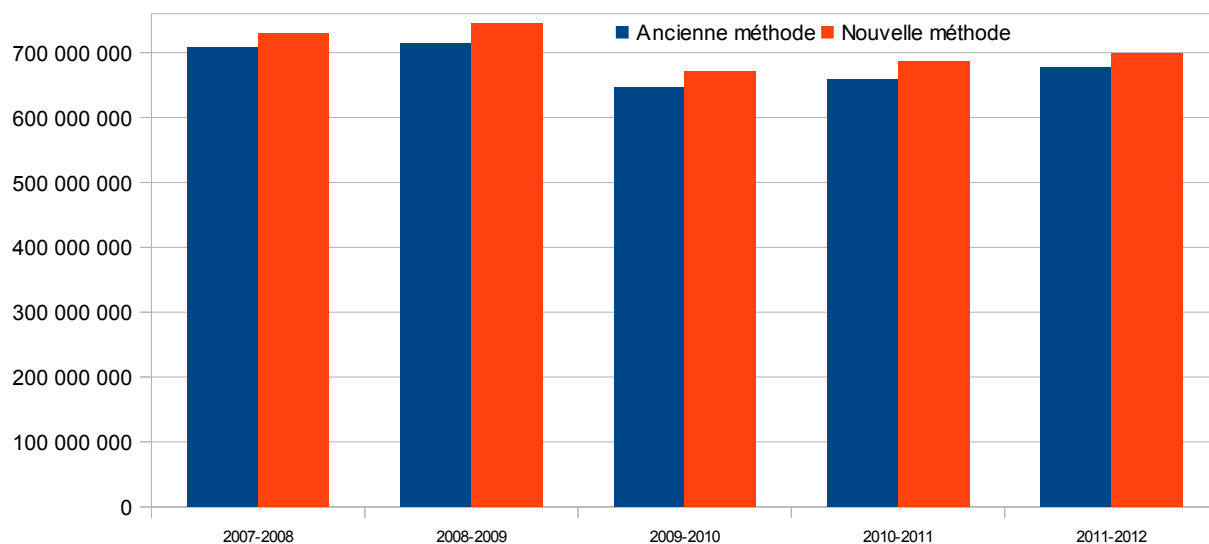


On constate que les intervalles de confiance sont particulièrement grands pour les *télésis à câble bas* et les *autres installations*, en raison des trop petits effectifs pour ces catégories, et que l'ancienne et la nouvelle méthode divergent très fortement sur cette catégorie (la nouvelle méthode donne des valeurs environ 40 % plus élevées que les valeurs trouvées avec l'ancienne).

Catégorie		1. Téléphériques bicâbles TBA, TBP et TBV	2. Télécabines DMD, DMV, TCD, TTCP, TMV et TMP	3.1 Télesièges fixes TSF	3.2 Télesièges débrayables TSCD et TSD	4.1 Téleskis fixes RFP	4.2 Téleskis débrayables RDP et RAE	5. Téleskis à câble bas RAC, RCOB et RCAB	6. Autres EDS, FUN et ASC
2008	Trafic mesuré	6 104 591	42 490 828	108 582 802	110 492 679	9 645 336	97 155 327	1 928 443	2 109 334
	Trafic estimé, nouvelle méthode	5 201 465	33 347 816	73 996 500	58 683 248	22 313 573	95 398 275	60 441 220	1 298 856
	Trafic total, nouvelle méthode	11 306 055	75 838 644	182 579 302	169 175 927	31 958 909	192 553 602	62 369 663	3 408 190
	Intervalle de confiance approximé à 95%	[11 122 348 ; 11 489 763]	[75 206 475 ; 76 470 813]	[182 036 351 ; 183 122 252]	[168 534 389 ; 169 817 465]	[31 715 415 ; 32 202 403]	[191 837 894 ; 193 269 309]	[59 803 966 ; 64 935 360]	[3 334 227 ; 3 482 153]
	Taille de l'intervalle de confiance	+/- 4 %	+/- 2 %	+/- 1 %	+/- 1 %	+/- 1 %	+/- 1 %	+/- 4 %	+/- 6 %
	Échantillon annuel	20	94	403	167	98	842	14	7
	Installations estimées	26	69	259	107	228	978	464	14
	Installations non estimées	0	0	0	0	1	2	20	0
2009	Trafic total, ancienne méthode	10 101 133	72 668 796	172 255 087	179 382 391	31 758 777	193 699 825	45 208 349	2 380 431
	Trafic mesuré	8 077 385	66 042 433	127 551 159	143 529 538	10 060 876	119 672 445	2 102 112	2 181 892
	Trafic estimé, nouvelle méthode	2 841 800	11 528 131	51 636 145	32 940 063	20 665 725	87 831 771	57 087 722	1 388 652
	Trafic total, nouvelle méthode	10 919 185	77 570 564	179 187 304	176 469 601	30 726 601	207 504 216	59 189 834	3 570 544
	Intervalle de confiance approximé à 95%	[10 658 615 ; 11 179 755]	[77 099 012 ; 78 042 115]	[178 873 526 ; 179 501 082]	[176 173 853 ; 176 765 349]	[30 568 061 ; 30 885 141]	[207 171 246 ; 207 837 187]	[56 601 758 ; 61 777 909]	[3 495 101 ; 3 645 986]
	Taille de l'intervalle de confiance	+/- 9 %	+/- 4 %	+/- 1 %	+/- 1 %	+/- 1 %	+/- 0,3 %	+/- 5 %	+/- 5 %
	Échantillon annuel	31	129	458	217	103	940	20	6
	Installations estimées	15	28	180	64	198	776	429	15
2010	Installations non estimées	0	0	0	0	1	1	12	0
	Trafic total, ancienne méthode	10 755 978	78 071 331	172 318 910	181 003 511	33 933 676	197 354 165	38 906 707	2 431 359
	Trafic mesuré	7 184 923	54 383 093	102 631 771	114 073 622	8 553 301	92 198 621	2 311 830	1 884 757
	Trafic estimé, nouvelle méthode	2 388 715	16 277 473	54 955 215	54 652 614	18 039 270	83 948 591	55 760 489	1 946 213
	Trafic total, nouvelle méthode	9 573 638	70 660 566	157 586 986	169 271 439	26 592 571	169 271 438	58 072 319	3 830 970
	Intervalle de confiance approximé à 95%	[9 391 421 ; 9 755 856]	[70 218 205 ; 71 102 926]	[157 154 668 ; 158 019 305]	[168 615 958 ; 169 926 920]	[26 494 169 ; 26 690 974]	[168 615 958 ; 169 926 919]	[54 679 411 ; 61 465 227]	[3 697 938 ; 3 964 003]
	Taille de l'intervalle de confiance	+/- 7 %	+/- 3 %	+/- 1 %	+/- 1 %	+/- 1 %	+/- 0,2 %	+/- 6 %	+/- 7 %
	Échantillon annuel	29	122	401	189	89	841	17	7
2011	Installations estimées	16	39	223	102	199	836	423	13
	Installations non estimées	0	0	0	0	1	1	13	0
	Trafic total, ancienne méthode	8 923 905	69 223 090	152 930 529	172 197 312	29 106 476	163 569 908	43 567 454	3 128 174
	Trafic mesuré	5 475 114	46 389 729	67 026 475	79 049 374	4 726 619	60 663 368	2 057 955	1 885 080
	Trafic estimé, nouvelle méthode	4 858 294	31 955 724	101 541 605	89 136 013	21 348 367	109 090 815	59 537 617	1 971 703
	Trafic total, nouvelle méthode	10 333 408	78 345 453	168 568 080	168 185 387	26 074 986	169 754 183	61 595 572	3 856 783
	Intervalle de confiance approximé à 95%	[10 158 686 ; 10 508 129]	[77 471 577 ; 79 219 329]	[168 177 762 ; 168 958 397]	[167 718 291 ; 168 652 482]	[25 866 373 ; 26 283 598]	[169 272 483 ; 170 235 884]	[58 174 423 ; 65 016 721]	[3 724 230 ; 3 989 336]
	Taille de l'intervalle de confiance	+/- 4 %	+/- 3 %	+/- 0,3 %	+/- 0,5 %	+/- 1 %	+/- 0,4 %	+/- 6 %	+/- 7 %
2012	Échantillon annuel	18	94	269	134	52	548	12	6
	Installations estimées	27	69	355	163	233	1 142	434	14
	Installations non estimées	0	0	0	0	1	0	20	0
	Trafic total, ancienne méthode	9 551 858	79 221 277	148 785 078	182 072 636	23 247 535	166 644 640	46 794 026	3 055 437
	Trafic mesuré	6 386 350	45 106 886	74 162 195	100 601 402	6 559 727	66 773 119	1 293 811	1 808 184
	Trafic estimé, nouvelle méthode	4 276 442	31 955 708	86 401 760	80 413 383	20 795 757	113 971 212	55 930 290	1 971 703
	Trafic total, nouvelle méthode	10 662 792	77 062 594	160 563 955	181 014 785	27 355 484	180 744 331	57 224 101	3 779 887
	Intervalle de confiance approximé à 95%	[10 304 085 ; 11 021 498]	[76 058 743 ; 78 066 445]	[160 144 179 ; 160 983 731]	[180 419 873 ; 181 609 697]	[27 235 313 ; 27 475 656]	[180 456 058 ; 181 032 604]	[53 896 575 ; 60 551 626]	[3 647 334 ; 3 912 440]
2012	Taille de l'intervalle de confiance	+/- 8 %	+/- 3 %	+/- 0,4 %	+/- 1 %	+/- 1 %	+/- 0,2 %	+/- 6 %	+/- 7 %
	Échantillon annuel	22	95	293	162	62	574	10	6
	Installations estimées	23	67	314	143	214	1 050	426	14
	Installations non estimées	1	0	0	0	1	0	14	0
	Trafic total, ancienne méthode	10 967 997	79 013 088	147 603 948	189 740 925	38 748 531	168 268 162	40 116 567	2 997 866

4.2.3 - Trafic total annuel

	2008	2009	2010	2011	2012
Trafic mesuré	378 509 340	479 217 840	383 221 918	267 273 714	302 691 674
Trafic estimé, nouvelle méthode	350 329 635	265 920 008	288 513 785	419 440 137	395 716 255
Trafic total, nouvelle méthode	729 190 292	745 137 848	671 735 703	686 713 851	698 407 929
Intervalle de confiance approximé à 95%	[726 506 081 ; 731 874 504]	[742 797 075 ; 747 478 621]	[668 579 174 ; 674 892 231]	[683 410 799 ; 690 016 903]	[695 196 118 ; 701 619 740]
Taille de l'intervalle de confiance	+/-1 %	+/-1 %	+/-1 %	+/-1 %	+/-1 %
Échantillon annuel	1 645	1 904	1 695	1 133	1 224
Installations estimées	2 145	1 705	1 851	2 437	2 251
Installations non estimées	23	14	15	21	16
Trafic total, ancienne méthode	707 454 789	714 775 639	646 997 968	659 372 485	677 457 084



Trafics annuels globaux

Les trafics globaux annuels calculés avec la nouvelle méthode sont de 3 à 4 % supérieurs à ceux estimés par l'ancienne. Les intervalles de confiance sur les trafics globaux sont très resserrés, les estimations sont donc d'une bonne fiabilité.

5 - Conclusion

La présente étude a été réalisée à partir des données annuelles mises à disposition par le STRMTG et DSF pour la période allant de 2008 à 2012.

5.1 - Synthèse des résultats de l'étude

Ce rapport présente les résultats obtenus à partir de méthodes de régression du trafic hivernal sur les remontées mécaniques de stations de montagne en fonction de variables explicatives, soit liées aux caractéristiques physiques des installations (capacité horaire, dénivelée, altitude inférieure...), soit liées au nombre de jours d'ouverture ou à la fréquentation mesurée par le nombre de journées-skieurs vendues dans la station.

La principale variable corrélée avec les trafics hivernaux est la capacité horaire. Elle explique une bonne partie de la variance des trafics et peut être utilisée comme unique variable explicative quantitative lorsqu'on ne dispose pas d'autres variables. Ce modèle « minimal » donne déjà des résultats corrects pour l'estimation des sommes de trafic par massif ou par catégorie. Il peut être enrichi par les caractéristiques physiques des installations qui sont significativement corrélées avec les trafics hivernaux :

- altitude inférieure et dénivelée pour les télécabines, télésièges et téléskis à attaches fixes ou débrayables ;
- altitude inférieure et longueur horizontale pour les téléphériques, téléskis à câble bas et autres installations.

Enfin, les données de fréquentation et de nombre de jours d'ouverture par remontée viennent également améliorer la précision des estimations individuelles, les jours d'ouverture par remontée étant la variable qui améliore le plus la qualité des estimations.

Finalement, des taux d'erreurs de 1 à 4 % sont constatés sur les sommes de trafic par massif et par catégorie, après évaluation par validation croisée sur les données dont on dispose.

En complément du présent rapport, un programme implémenté sous R (logiciel de statistiques libre et gratuit) est fourni. Ce programme permet de mettre en œuvre la méthode présentée dans ce rapport sur des données annuelles de trafic. Le programme a été testé sur les données de 2008 à 2012 fournies par le STRMTG et par DSF. Il permet d'une part de fournir des estimations de trafic agrégées par massif ou par catégorie ainsi qu'une estimation du trafic global sur l'ensemble du parc de remontées mécaniques pris en compte. D'autre part, il fournit des éléments d'évaluation des modèles utilisés et de diagnostic, éléments mis à jour automatiquement avec chaque nouveau jeu de données annuel.

5.2 - Précaution d'utilisation des chiffres

La précision des estimations de trafic croisant les catégories et des massifs n'a pas été évaluée dans le cadre du présent travail et pourrait être faible en raison des petits échantillons de certaines classes. Leur éventuelle utilisation n'est pas recommandée.

5.3 - Pistes d'amélioration

L'examen des données suggère des possibles pistes d'amélioration de la précision des estimations en jouant sur quatre plans :

1. Il est souhaitable d'augmenter la taille et la diversité des échantillons de trafic recueillis. Pour cela, on pourrait envisager de programmer des relances auprès des exploitants en visant certains équipements particuliers, notamment les *téléphériques bicâbles*, les *téléskis à câble bas*, les *autres installations* ainsi que les tapis, non pris en compte pour le moment. Des relances pourraient également être mises en œuvre auprès de certains non-répondants systématiques : à partir de la base de données des trafics 2008-2012, 169 non-répondants systématiques peuvent être identifiés parmi les exploitants. Une analyse plus approfondie de cette liste permettrait d'identifier une liste de stations prioritaires pour améliorer le recueil de données (notamment en complétant les catégories de remontées mécaniques et les massifs les moins bien représentés).
2. Il serait nécessaire de compléter la base de données FIRM en renseignant les valeurs manquantes sur certaines variables, en particulier la capacité (débit maximum horaire montant), l'altitude inférieure et la dénivelée. Les codes « FIRM » ne sont pas tous attribués aux remontées dans les données fournies par DSF de 2008 à 2012, empêchant leur utilisation pour estimer les trafics puisqu'il n'est pas possible de faire une jointure automatique entre les deux sources de données (FIRM et données DSF). Une uniformisation et une maintenance commune des deux bases améliorerait certainement la qualité des estimations réalisées.
3. Une fois les deux premiers points réalisés, un travail de recalibrage et une reconsidération des variables explicatives du trafic hivernal pourrait être réalisé pour améliorer la précision des estimations fournies à ce stade.
4. L'estimation n'est réalisée actuellement que sur la saison hivernale, la plupart des données de trafic n'étant remplie que pour l'hiver. Afin d'évaluer un trafic annuel, il serait souhaitable d'inciter les exploitants à remplir également le trafic et le nombre de jours d'ouverture (voire le nombre de titres vendus) de l'été.

Rédigé, le 09 août 2013 à Bron

La chargée d'Affaire

Aurélie BOUSQUET

Vu et vérifié, le 14/08/2013

Le responsable de l'unité

Olivier Richard

Références

- [1] P. Lebreton, « Les indices nationaux de la circulation routière Pourquoi ? Pour qui ? Comment ? », SETRA, 7, mars 1986.
- [2] P. Cousin, D. Schmitt, et D. Postel, « Nouvel indice de circulation sur le réseau routier national, résultats de 2008. », SETRA, Note d'information n°3, mai 2009.
- [3] R. Rakotomalala, « Econométrie : la régression linéaire simple et multiple », Université Lyon 2, Cours, févr. 2013.
- [4] R. Rakotomalala, « Pratique de la Régression Linéaire Multiple - Diagnostic et sélection de variables », Université Lyon 2, Cours, juill. 2011.
- [5] Y. Dodge, *Analyse de régression appliquée*. Paris: Dunod, 2004.
- [6] J. Jacques, « Modélisation Statistique », Université de Lille 1, Cours.
- [7] W. G. Hopkins, « New View of Statistics: Models - Details », 2001. [En ligne]. Disponible sur: <http://www.sportsci.org/resource/stats/modelsdetail.html>. [Consulté le: 07-août-2013].
- [8] A. Charpentier, « Confidence interval for predictions with GLMs | Freakonometrics », nov-2011. [En ligne]. Disponible sur: <http://freakonometrics.hypotheses.org/2289>. [Consulté le: 07-août-2013].
- [9] A. Charpentier, « sur les transformations dans un modèle linéaire | Freakonometrics », nov-2012. [En ligne]. Disponible sur: <http://freakonometrics.hypotheses.org/126>. [Consulté le: 07-août-2013].
- [10] J. M. Ver Hoef et P. L. Boveng, « Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? », *Ecology*, vol. 88, n° 11, p. 2766-2772, nov. 2007.
- [11] R. Lancelot et M. Lesnoff, « Sélection de modèles avec l'AIC et critères d'information dérivés », nov-2005. [En ligne]. Disponible sur: ftp://ftp.cirad.fr/pub/group-r/groupe-r/Fiches/AIC_v3.pdf. [Consulté le: 07-août-2013].

Annexes

Annexe A – Éléments de méthodologie sur les régressions

Les techniques de régression permettent de modéliser des relations complexes entre une variable à estimer et des variables explicatives. On peut les utiliser pour l'estimation de valeurs manquantes pour la variable à expliquer, toutefois leur mise en œuvre est un peu plus lourde que celle des méthodes de redressement par un ratio. De plus, les méthodes de régression nécessitent des tailles d'échantillon assez importantes, d'autant plus lorsqu'on augmente le nombre de variables explicatives, puisqu'on exige généralement un nombre d'observations représentant 10 à 20 fois le nombre de variables explicatives pour caler le modèle.

Modèle linéaire

Le modèle linéaire est le type de régression le plus communément utilisé pour modéliser le lien qui existe entre une variable à expliquer (ici le trafic) et une ou des variables explicatives à déterminer. Le modèle linéaire est basé sur un ajustement au sens des moindres carrés ordinaires. Il suppose l'existence d'une relation linéaire entre la variable à expliquer et une série de variables explicatives. Parmi ses hypothèses figurent la normalité de la variable à expliquer et des variables explicatives ainsi que la constance de la variance de la variable à expliquer sur l'ensemble du domaine de valeurs. Une présentation détaillée du modèle linéaire peut être trouvée dans [3], [4], [5] et [6]. Nous revenons ici sur les principaux points utiles pour cette étude.

Forme de la variable à expliquer

Dans le modèle linéaire, on suppose que la variable à expliquer Y suit une loi normale et que sa variance est indépendante de sa moyenne (autrement dit l'écart à la moyenne est le même que l'on soit dans les faibles ou dans les fortes valeurs de Y). Lorsque ce n'est pas le cas, on cherche à se rapprocher au maximum de ces hypothèses par des transformations simples de la variable à expliquer (ex : $\log(Y)$, $1/Y$, $\sqrt{Y}$)... Les transformations qui normalisent une distribution ont généralement également la propriété de rendre la variance plus uniforme et vice et versa.

La transformation en logarithme est classiquement utilisée lorsque la variance est proportionnelle au carré de la moyenne (c'est-à-dire que l'écart-type est proportionnel à la moyenne). La transformation en racine carrée est classiquement utilisée lorsque la variance est proportionnelle à la moyenne, comme c'est le cas de données distribuées selon une loi de Poisson, distribution couramment rencontrée pour les données de comptage.

Le test de Shapiro-Wilk permet de tester la normalité de la variable Y ou de ses transformées. Lorsque la p-valeur du test est inférieure au seuil fixé, on peut rejeter l'hypothèse nulle selon laquelle la variable suit une loi normale. Plus la p-valeur est élevée, plus la probabilité que la distribution soit compatible avec une loi normale augmente. On ne cherchera pas nécessairement une p-valeur > 0.05 (seuil de 5 % utilisé classiquement) mais plutôt la plus grande p-valeur possible. En effet, rendre la variable totalement compatible avec la loi normale est généralement illusoire sur des données réelles et on se contente donc souvent d'une transformation rendant la variable « la plus normale possible ».

Choix des variables explicatives

Les variables explicatives peuvent être quantitatives ou qualitatives. Lorsqu'elles sont qualitatives, on procède généralement à un codage pour les insérer dans le modèle linéaire au même titre que les variables quantitatives. Le codage le plus simple est un codage sous forme d'indicateurs : chaque modalité de la variable qualitative fait l'objet d'une nouvelle variable valant 1 lorsque la modalité est rencontrée, 0 sinon. Dans cette étude, on adopte un tel codage pour la variable *massif*.

Lors de l'insertion de variables explicatives supplémentaires dans un modèle linéaire, on s'assure que les colinéarités entre l'ensemble des variables explicatives sont limitées, afin d'obtenir un modèle fiable. Une trop forte colinéarité entre les variables explicatives peut en effet produire des résultats très instables (une petite modification des observations sur les variables explicatives provoque une forte modification des paramètres de la régression), les signes et les valeurs des paramètres peuvent devenir incohérents par rapport aux autres variables et aux connaissances qu'on a sur le domaine, certaines variables explicatives peuvent alors paraître non pertinentes dans le modèle, uniquement à cause d'une variance mal estimée, et ainsi éliminées malgré leur forte corrélation avec la variable à expliquer.

Diagnostic du modèle

Le test de significativité de Fisher sur l'ensemble du modèle permet de déterminer si la régression linéaire a un sens par rapport aux données observées. Une p-valeur inférieure au seuil de significativité (pris dans cette étude à 5 %) indique que le modèle est significatif. Des tests de significativité peuvent également être réalisés par paramètre du modèle. Les variables dont le paramètre ajusté est non-significatif pourront être retirées du modèle.

Le coefficient de détermination, noté R^2 , nous renseigne sur le pouvoir explicatif du modèle : il correspond au pourcentage de la variance de Y expliquée par le modèle.

Enfin, l'analyse des résidus du modèle linéaire permet de détecter d'éventuelles violations des hypothèses du modèle. Les résidus d'un « bon » modèle doivent avoir les propriétés suivantes :

- distribution normale uniforme (remarque : la normalité des résidus est un objectif, on considère en pratique que le fait d'avoir une distribution « d'allure normale » est suffisant pour que le modèle puisse être utilisé en prévision, l'allure normale se résumant à une distribution globalement symétrique et en forme de cloche, voir [7]) ;

- homéoscédasticité : constance de la distribution des résidus sur l'ensemble de la plage de valeurs des Y : si leur moyenne varie, cela signifie que les variables explicatives ne sont pas adaptées (manque un terme, pas de la bonne forme...), si leur écart-type varie, cela signifie que la variable à expliquer n'a pas la bonne forme et qu'il faut utiliser une transformation
- indépendance : il ne doit pas apparaître de corrélation entre les résidus.

Transformation inverse de la variable à expliquer

Lorsque le modèle linéaire a été ajusté sur une variable transformée, comme c'est le cas ici, il faut déterminer la transformation inverse à effectuer pour retomber sur des prévisions et un intervalle de confiance pour Y et non pour la transformée de Y.

[9] nous donne les transformations à effectuer pour le cas d'une transformation en racine carrée. Si on note $Z = \sqrt{Y}$, $\hat{Y}$ la prévision pour Y et $\hat{Z}$ la prévision pour Z, on a :

$$\hat{Y} = \hat{Z}^2 + \text{Var}(Z)$$

$$IC(\hat{Y})_{\alpha} = [(\hat{Z} - \text{Var}(Z) * \sqrt{\chi^2_{1-\alpha,1}})^2; (\hat{Z} + \text{Var}(Z) * \sqrt{\chi^2_{1-\alpha,1}})^2], \text{ où } \chi^2_{1-\alpha,1} \text{ est le quantile } 1-\alpha \text{ de la loi du } \chi^2 \text{ à 1 degré de liberté.}$$

Dans certains cas, une transformation de la variable à expliquer est suffisante pour linéariser la relation entre Y les variables explicatives et aboutit à des erreurs distribuées normalement et de variance homogène. Dans d'autres cas, la simple transformation ne permettra pas de résoudre le problème et l'utilisation d'un *modèle linéaire généralisé* sera nécessaire. Nous présentons ici les modèles de Poisson, quasi-poissonien et négatif binomial, particulièrement adaptés aux données de comptage.

Modèle linéaire généralisé

Les données de comptages sont classiquement décrites par une loi dite « de Poisson ». Comme cela a été dit ci-dessus, le modèle linéaire n'est généralement pas adapté à de telles données de par ses hypothèses (résidus suivant une loi normale, variance homogène). Son utilisation sur des données de comptage peut parfois conduire à réaliser des estimations de faible précision. De plus, la nature même des données de comptage (discrètes, positives, asymétriques avec une forte proportion de faibles valeurs) peut rendre les moindres carrés ordinaires inopérants.

Le modèle linéaire généralisé peut alors être utilisé à la place du modèle linéaire. On s'appuie alors sur le principe du maximum de vraisemblance, la vraisemblance d'un modèle étant la probabilité d'observer les données que l'on a avec ce modèle.

Modèle de Poisson

Lorsque l'on suppose que la loi sous-jacente de la variable à expliquer est une loi de Poisson, on utilise la régression de Poisson qui s'appuie sur le formalisme du modèle linéaire généralisé. Le modèle de Poisson suppose l'égalité de la moyenne et de la variance.

Traitement de la sur-dispersion des données

L'hypothèse très restrictive d'égalité de la moyenne et de la variance est rarement vérifiée sur des données de comptage expérimentales. C'est pourquoi des distributions permettant de prendre en compte une sur-dispersion (variance supérieure la moyenne) sont également souvent utilisés pour modéliser des données de comptages :

- le modèle quasi-poissonien, qui est un modèle de Poisson dans lequel on utilise une notion légèrement modifiée de la vraisemblance et qui représente la variance comme une fonction linéaire de la moyenne ;
- le modèle négatif binomial, qui suppose que la variance est une fonction quadratique de la moyenne.

Dans ces deux modèles, la variance étant une fonction différente de la moyenne, les trafics faibles ou forts sont pondérés différemment. Le choix entre ces deux modèles est délicat a priori et on teste généralement les deux afin de comparer les erreurs qu'ils génèrent sur les données dont on dispose [10].

Caractéristiques des résidus

Dans un modèle linéaire généralisé, les résidus ne sont plus normalement distribués. La transformation d'Anscombe permet toutefois de retrouver des résidus dont, lorsque le modèle est bon, la distribution est normale.

Test d'ajustement

Pour les modèles de Poisson, quasi-Poisson et négatif binomial on utilise un test d'ajustement du χ^2 : le modèle est considéré comme significatif si $1 - \frac{\text{deviance}(m)}{df(m)}$ est supérieure au seuil de 0,05 (m est le modèle et df le nombre de degrés de liberté résiduels).

Intervalle de confiance de la prévision

Le formalisme des modèles linéaires généralisés utilise une fonction de lien qui correspond à une transformation de la variable à expliquer. Dans le cas des modèles poissonien, quasi-poissonien et négatif binomial, c'est une fonction de lien logarithmique. Nous allons donc travailler sur une transformation qu'il faudra inverser. Les logiciels de statistiques donnent généralement directement la valeur de la prévision après transformée inverse ainsi que sa variance⁷, ce qui nous permet de calculer un intervalle de confiance de la prévision de façon très classique à partir de la variance et du quantile d'une loi gaussienne, comme indiqué dans [8].

$$IC(\hat{Y})_{\alpha} = [\hat{Y} - \sqrt{\text{Var}(\hat{Y})} * z_{1-\frac{\alpha}{2}}; \hat{Y} + \sqrt{\text{Var}(\hat{Y})} * z_{1-\frac{\alpha}{2}}]$$

Comparaisons de modèles

Comparaison de modèles imbriqués à l'intérieur d'une même famille

⁷ Dans R, c'est la fonction glm qui ajuste le modèle linéaire généralisé et la fonction predict avec l'option « response » qui nous donne la prédiction et la variance pour la variable après transformée inverse.

Les critères AIC et BIC⁸ sont couramment utilisés pour évaluer la qualité d'ajustement de modèles linéaires généralisés. Ils s'appliquent également aux modèles linéaires classiques. Toutefois, les indicateurs AIC et BIC sont conçus pour comparer des modèles d'une même famille, évalués sur un même ensemble de données et emboîtés (c'est-à-dire qu'ils dérivent tous d'un même modèle complet par suppression de certaines variables) [11].

Pour la sélection de variables, on utilise des méthodes de sélection *ascendante* (on ajoute une à une des variables dans le modèle), *descendante* (on supprime une à une des variables du modèle) ou *pas à pas* (on ajoute ou on supprime une variable à chaque itération). Le choix des variables à ajouter ou supprimer se fait sur la base de l'indicateur AIC, le but étant de le faire diminuer à chaque itération. La procédure s'arrête quand on ne peut plus le faire diminuer.

Comparaison de modèles de familles différentes

Pour comparer les qualités prédictives de modèles issus de différentes familles, on utilise couramment une procédure de validation croisée. La validation croisée consiste à découper aléatoirement l'échantillon de données en K sous-ensembles, puis à utiliser K-1 de ces sous-ensembles pour caler la régression et le K^{ème} pour évaluer l'erreur commise par le modèle. On peut ainsi répéter K fois la procédure en changeant le sous-ensemble mis à part et faire la moyenne des K erreurs obtenues. On prend $K = E(\sqrt{n})$, où n est la taille de l'échantillon et E(x) désigne la partie entière de x.

Pour des données de comptage, trois mesures de l'erreur commise par les modèles peuvent être utilisées.

Racine carrée de l'erreur quadratique moyenne

$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$, où $\hat{Y}_i$ représente la prévision de trafic et Y_i la valeur observée.

Cet indicateur est également sensible aux erreurs positives et aux erreurs négatives et dans le champ des petites et des grandes valeurs. C'est la mesure d'erreur la plus couramment utilisée.

Racine carrée de l'erreur quadratique moyenne du logarithme

$$RMSE.Log = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(Y_i) - \log(\hat{Y}_i))^2}$$

Lorsqu'on travaille avec des données de comptage, la sensibilité aux erreurs n'est pas la même dans les petites et les grandes valeurs et il est peu probable que les erreurs soient symétriques distribuées autour de 0. Se tromper de 10 sur une prévision de 1000 est moins grave que de se tromper de 10 sur une prévision de 100.

8 Ces indicateurs sont définis de la manière suivante : AIC = $-\log(L) + k$ et BIC = $-\log(L) + k \cdot \log(N)$, où L est la vraisemblance du modèle, N est le nombre de données d'apprentissage, k est une constante généralement prise égale à 2.

Cette mesure d'erreur prend en compte cet aspect au travers du logarithme. Elle est particulièrement pertinente lorsque le modèle linéaire généralisé a été estimé avec une fonction de lien logarithmique.

Erreur du χ^2

$$\chi^2 = \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\hat{Y}_i}$$

Une troisième méthode consiste à utiliser l'erreur du χ^2 . De la même manière que l'erreur quadratique du logarithme moyen, elle tient compte de l'échelle de valeur des prévisions, en revanche, elle tient compte de manière symétrique des erreurs positives et négatives. Elle s'interprète comme un pourcentage d'erreur moyen par rapport à la valeur prédite.

Enfin, nous utilisons deux indicateurs d'erreurs adaptés à notre problème : estimer des sommes de trafics sur un sous-ensemble d'installations (par catégorie ou par massif).

Erreur relative sur la somme

Cette erreur peut se lire comme un pourcentage d'écart à la valeur prédite.

$$\Sigma = \frac{\sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i}{\sum_{i=1}^n \hat{Y}_i}$$

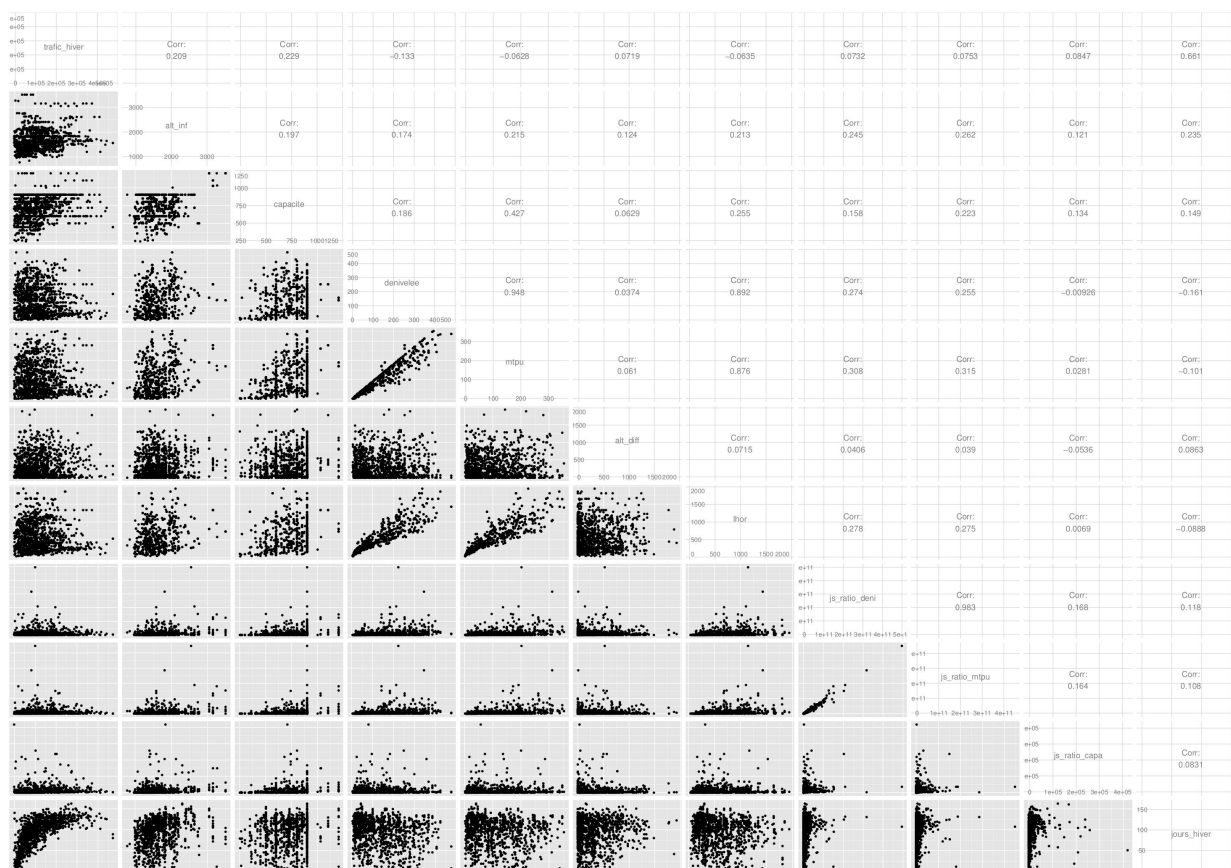
Erreur du χ^2 sur la somme

C'est l'homologue de l'erreur du χ^2 en considérant que la variable d'intérêt est la somme des trafics.

$$\Sigma \chi^2 = \frac{(\sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i)^2}{\sum_{i=1}^n \hat{Y}_i}$$

Annexe B – Graphiques de corrélation des variables explicatives et de la variable à expliquer

Les graphiques ci-dessous présentent des appariements de variables qui permettent de détecter les corrélations qui existent entre ces variables et la nature de ces corrélations (linéaires ou non). De tels graphiques sont générés par l'outil sous R pour toutes les catégories et pour toutes les périodes souhaitées (annuelle ou pluri-annuelle). Le présent graphique présente les résultats pour la catégorie 4 des téléskis à perches fixes ou débrayables sur l'ensemble de la période de 5 ans.



Appariement de variables pour la catégorie 4 (téléskis à perches fixes ou débrayables) sur la période 2008-2012

Annexe C – Traitements réalisés sur les données

Les données étant parfois incomplètes ou incertaines, des traitements ont été réalisés en préalable aux analyses statistiques.

Données de trafic et de nombre de jours d'ouverture par remontée

Il n'était pas possible de faire la différence de façon certaine entre les trafics recensés à 0 (remontée fermée cette année-là) et l'absence de données. Toutes les valeurs mises à 0 ont par conséquent été considérées comme des données absentes.

Code station

Le code de la station est utilisé pour calculer les ratios de dénivelée, de moment de puissance et de capacité de la remontée par rapport à l'ensemble de la station et répartir ainsi les journées-skieurs sur l'ensemble des remontées. Sur certaines remontées, la station n'était pas renseignée. Lorsque cela était possible, les codes de station manquants ont été reconstitués à partir du nom de la station. Pour certaines installations, le nom n'était pas renseigné et la reconstitution n'a donc pas été possible.

Jours d'ouverture hivernale

Les valeurs de jours d'ouverture hivernale supérieures à 200 ont été supprimées.

Journées-skieurs vendues

Les journées-skieurs recensées à 0 ont été supprimées de la base. On les considère comme des valeurs manquantes.

Problème d'affectation d'une remontée à une catégorie en 2010

En 2010, la remontée d'identifiant 380 655 a été affectée à la catégorie « RCAB ». Pourtant, avec ses 568 mètres de dénivelée, il s'agit vraisemblablement d'une erreur dans la base. Ceci est confirmé par les données de 2011, l'identifiant 380 655 correspondant alors au télésiège de l'Oursière (catégorie TSD). Pour ne pas fausser les résultats du modèle sur 2010, la remontée a donc été affectée à la catégorie TSD.

Annexe D – Catégories d'installations de remontées mécaniques

<i>Libellé</i>	<i>Nom</i>	<i>Catégories supposées à comportements homogènes pour l'estimation des trafics</i>	<i>Catégories utilisées pour les agrégations de trafic annuelles</i>
TBP	Téléphériques bicâbles pulsés	1 – téléphériques bicâbles	1 – téléphériques bicâbles
TBV	Téléphériques bicâbles à va et vient		
TBA	Autres types de téléphériques bicâbles		
DMD	Double monocâbles à attaches débrayables	2 – télécabines ou téléphériques monocâbles	2 – télécabines ou téléphériques monocâbles
DMV	Double monocâbles à va et vient		
TCD	Télécabines à attaches débrayables		
TCP	Télécabines pulsées		
TMV	Téléphériques monocâbles à va et vient		
TPM	Autres types de téléphériques monocâbles		
TSD	Télesièges à attaches débrayables	3 – télesièges	31 – télesièges débrayables
TSCD	Téléphériques monocâbles avec sièges + cabines		
TSF	Télesièges à attaches fixes		32 – télesièges fixes
RDP	Téléskis à perches débrayables	4 – téléskis	41 – téléskis débrayables
RAE	Téléskis à enrouleurs		
RFP	Téléskis à perches fixes		42 – téléskis fixes
RCAB	Téléskis à câble bas	5 – téléskis à câble bas	5 – téléskis à câble bas
RCOB	Téléskis à corde bas		
RAC	Téléskis de type "télécorde"		
ASC	Ascenseurs inclinés	6 – autres	6 – autres
FUN	Funiculaires		
EDS	Engins divers		
TRSM (non pris en compte)	Tapis	7 – tapis	7 – tapis