
MISE EN PLACE D'UNE ENQUÊTE AUPRÈS DES VOYAGEURS EN CORSE

Aude GENOVESE-BOLLEYN (*), Thomas MERLY-ALPA (**), Cyril FAVRE-MARTINOZ (***)

(*) Insee, Direction régionale de Corse

(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

(***) Insee, Direction régionale de La Réunion Mayotte

aude.genovese-bolleyn@insee.fr

cyril.favre-martinoz@insee.fr

thomas.merly-alpa@insee.fr

Mots-clés : Enquêtes sur le tourisme, échantillonnage, collecte, équilibrage.

Résumé

Cette communication vise à présenter les objectifs et le protocole de l'enquête auprès des voyageurs en Corse. Elle présente également les spécificités du plan de sondage mis en œuvre pour obtenir l'échantillon qui prenait en compte des contraintes de collecte liés à la dispersion géographique des lieux de collecte et à l'activité des enquêteurs tout en optimisant la précision statistique via des méthodes de sondage équilibré.

L'enquête auprès des voyageurs en Corse, menée entre mai et novembre 2017, s'inscrit dans l'action de l'Agence du Tourisme de la Corse (ATC) permettant d'éclairer les politiques publiques concernant l'économie touristique de l'île. Cette opération a pour objectif de mesurer le tourisme en Corse à travers la connaissance de la clientèle (profil des visiteurs, nationalités des touristes venant en Corse, hébergement marchand ou non marchand), les canaux de réservation, la répartition des séjours sur le territoire. L'objectif de l'enquête est de quantifier la répartition des nuitées passées en Corse par les touristes selon les différents type d'hébergement marchands (hôtels, campings, AHCT, gîtes, locations entre particuliers...) et non marchands (famille ou amis, séjours du propriétaire dans sa résidence secondaire.) Il s'agit aussi de mieux connaître les différentes clientèles et leurs pratiques selon les périodes : avant saison, cœur de saison et après saison, vacances scolaires ou non. Les principales variables d'intérêt pour la population touristique sont l'origine géographique, le motif de voyage, le mode d'hébergement, la durée du séjour, le lieu d'hébergement, le degré de satisfaction, le montant des principales dépenses. La problématique de l'étalement de la saison touristique, très centrée sur les mois de juillet et d'août est au cœur des préoccupations de l'ATC et des politiques publiques concernant l'économie touristique de l'île.

L'enquête 2017 est réalisée auprès des passagers (résidents en Corse ou touristes) quittant la Corse via ses cinq principaux ports (Ajaccio, Bastia, l'île Rousse, Porto-Vecchio et Bonifacio) et ses quatre aéroports (Ajaccio, Bastia, Figari et Calvi), ayant passé au moins une nuitée sur la Corse. L'unité sur laquelle porte l'enquête est l'« entité budgétaire », c'est-à-dire le groupe de personnes voyageant ensemble et partageant les dépenses. Pour les voyages en bateaux, les unités primaires de sondage sont directement des traversées au départ de l'île. Pour les voyages en avion, les unités primaires sont des plages horaires de 2 heures, l'ensemble des vols décollant à l'intérieur de ces plages étant alors enquêtés.

Pour éviter des collectes simultanées dans une même zone, il est nécessaire que le recouvrement entre les deux échantillons soit faible voire inexistant. Afin d'éviter les recouvrements entre l'échantillon de tranches horaires d'avions et les départs des bateaux, un tirage équilibré à probabilités égales des jours dédiés aux avions et des jours dédiés aux bateaux a été réalisé au sein de chaque strate.

À partir des jours tirés précédemment, on construit une base restreinte des tranches horaires correspondant à l'ensemble des tranches horaires et/ou de bateaux appartenant à un des jours tirés. Ensuite, au sein de chaque strate, les probabilités d'inclusion des tranches horaires et des bateaux sont définies proportionnellement au nombre de passagers (en utilisant des taux de remplissage moyens). À partir de ces probabilités d'inclusion, on réalise un tirage équilibré au sein de chaque strate mobilisant des variables telles que :

- le nombre de passagers
- les indicatrices de jours de la semaine
- l'indicatrice de jour de vacances
- le nombre de passagers transportés selon le type de compagnie
-

La collecte de l'échantillon a été faite sous la forme d'un questionnaire auto-administré. Composé de deux pages, le questionnaire disponible en plusieurs langues (français, anglais, italien et allemand) et auto-administré (sous format papier) est distribué exhaustivement par un enquêteur aux groupes de passagers aériens après les contrôles de sécurité dans la salle d'embarquement et aux passagers maritimes dans la salle d'embarquement et dans les files d'attente des véhicules à l'embarquement (un questionnaire par entité budgétaire). La récupération des questionnaires renseignés s'effectue par les enquêteurs au moment de l'embarquement proprement dit dans l'avion ou dans le bateau.

Des opérations de redressement ont été menées à partir des données de la collecte, pour l'apurement, la correction de la non-réponse puis le calage ; ces opérations ont conduit à se poser des questions sur les marges à retenir, en particulier sur la notion de résidants et de touristes, ainsi que vis-à-vis des autres sources statistiques sur le tourisme.

1. Présentation de l'enquête

1.1. Objectifs

L'enquête statistique « auprès des voyageurs au départ de la Corse » répond à un fort besoin local visant à mieux connaître les parts de marché des différents modes d'hébergement, notamment celles concernant les hébergements non marchands ou les locations saisonnières. Ainsi, il s'agit de mieux connaître les différentes clientèles et leurs pratiques selon les périodes : avant saison, cœur de saison et après saison.

Les principales variables d'intérêt pour la population touristique sont l'origine géographique, la préparation du voyage, le mode d'hébergement, la durée du séjour, le lieu principal d'hébergement, le degré de satisfaction et la structure des dépenses.

La problématique de l'étalement de la saison touristique, très centrée sur les mois de juillet et août, est au cœur des préoccupations de l'Agence du Tourisme de la Corse (ATC) et des politiques publiques concernant l'économie touristique de l'île.

1.2. Contexte : éléments existants

Les enquêtes de fréquentation touristique de l'INSEE (Hôtels – Campings – Autres Hébergements collectifs) dénombrent les nuitées passées dans ces types d'hébergement collectifs marchands. L'utilisation des résidences secondaires et les locations saisonnières entre particuliers notamment sont des sujets très prégnants et peu quantifiés .

Il n'existe pas, pour la Corse, de données alternatives portant sur les thèmes de l'enquête. Les enquêtes SDT (Suivi de la Demande Touristique) et EVE (Enquête auprès des Visiteurs Étrangers en France) de la DGE ne fournissent pas de résultats régionaux correspondant aux objectifs de l'enquête.

L'Observatoire des Transports de la Corse enregistre les variations journalières de la population présente sur l'île en décomptant, au jour le jour, les arrivées et départs des passagers dans toutes les infrastructures portuaires et aéroportuaires de l'île.

L'enquête auprès des voyageurs apporte donc une information complémentaire nécessaire à l'analyse du tourisme en Corse ; c'est pourquoi elle a obtenu le label d'intérêt général et de qualité statistique du CNIS, émis le 23 janvier 2017.

1.3. Un partenariat INSEE-ATC

L'enquête est réalisée en partenariat entre la direction régionale de l'Insee de Corse et l'Agence du Tourisme de la Corse (ATC), agissant pour le compte de la Collectivité unique de Corse.

La demande émane de l'Agence du Tourisme de la Corse (ATC) qui agit, en matière de tourisme, pour le compte de la Collectivité Unique de Corse. Les CCI, parties prenantes de l'opération, font aussi partie des principaux utilisateurs potentiels des résultats.

Le montage partenarial de cette opération repose sur le fait que l'Insee est responsable de la conception statistique de l'enquête, de son exploitation et de son suivi. L'agence du tourisme de la Corse prend en charge l'ensemble des coûts externes, la gestion des enquêteurs et les relations avec les autorités portuaires et aéroportuaires.

L'Insee intervient en tant que maître d'ouvrage sur la méthodologie et les traitements statistiques. Il assure :

- la conception méthodologique de l'enquête ainsi que son exploitation (extrapolation, redressements, traitements de la non-réponse, tabulation des résultats),
- le suivi et le pilotage de la collecte et le suivi des enquêteurs avec l'ATC.

1.4. Un partenariat DR-DG

La DR de Corse a constitué une base de sondage constituée des plannings prévisionnels des compagnies par infrastructures et a estimé les passagers transportés.

A partir de cette base, le département de la méthodologie statistique a effectué le tirage de l'échantillon, décrit en partie 3.1. Une formation et un appui technique de la part de la DMCSI a permis à la DR d'effectuer les redressements, décrits en partie 3.2. Le questionnaire a été élaboré par l'Insee Corse et l'ATC. La DMCSI, la DDAR et la DGE ont contribué à l'améliorer.

2. Déroulement de l'enquête

2.1. Champ, période et cible de l'enquête

Le champ de l'enquête est l'ensemble des passagers (résidents en Corse ou touristes) quittant la Corse via ses principaux ports (Ajaccio, Bastia, l'Île Rousse, Porto-Vecchio et Bonifacio) et ses quatre aéroports (Ajaccio, Bastia, Figari et Calvi).

La période d'observation s'étale du 1er mai 2017 au 5 novembre 2017.

L'unité de collecte est l'« entité budgétaire », c'est-à-dire le groupe de personnes voyageant ensemble et partageant les dépenses. Cette notion conduit à interroger la personne qui gère les dépenses pour le groupe enquêté. En effet, cette personne est la plus à même de connaître les dépenses réalisées pendant le séjour et de décrire le comportement du groupe en matière de tourisme.

Cette notion est aussi celle utilisée dans les enquêtes SDT et EVE de la DGE.

2.2. Questionnaire

Le questionnaire anonyme et auto-administré, sous format papier (A4, recto-verso), est distribué exhaustivement par un enquêteur aux groupes de passagers aériens. Pour garantir un taux de réponse optimal et garder le répondant attentif, le questionnaire papier comporte 24 Questions.

Les 13 premières questions ainsi que la dernière permettent de caractériser le voyageur, et de déterminer son niveau de fidélisation avec la Corse ; une question vise à mieux situer les principales destinations concurrentes à la Corse.

Les quatre questions suivantes permettent de caractériser l'organisation du voyage, puis quatre autres questions portent sur l'hébergement, deux autres questions apportent des précisions sur le déroulement du séjour et enfin deux questions traitent de la satisfaction et de l'intention de retour.

Le questionnaire complet est en Annexe 1 de l'article.

2.3. Mode de collecte

La remise des questionnaires est autant que possible exhaustive pour un avion ou un bateau. Des stylos et planchettes sont proposés aux groupes enquêtés afin de faciliter leur réponse. L'importance de la présentation de l'enquête par l'enquêteur est primordiale pour améliorer le taux de réponse des enquêtés, notamment celui des résidents en Corse. Pour ces derniers, les enquêteurs, autant qu'ils le peuvent, posent directement les 6 questions les concernant, et remplissent les questionnaires.

La récupération des questionnaires renseignés s'effectue par les enquêteurs au moment de l'embarquement dans l'avion ou dans le bateau.

Le questionnaire est disponible en plusieurs langues : français, anglais, italien et allemand.

Les CCI ont délivré les autorisations d'accès (badges avec photos) aux enquêteurs. Toutefois, il a fallu prévoir, plus particulièrement pour les aéroports, un délai de 4 semaines pour la délivrance de ces badges.

Les protocoles de collecte des ports et des aéroports différent :

- Pour les ports : deux flux devaient être traités : les véhicules et les piétons. Trois enquêteurs ont été nécessaires, deux pour le flux de véhicules et un dans la salle d'embarquement des piétons. Le flux pour les véhicules a pu être traité de façon isolée. En effet les couloirs de circulation et de files d'attente sont séparés selon les départs de bateaux. La remise des questionnaires a eu lieu après le contrôle de remise des documents d'embarquement, lorsque les véhicules étaient à l'arrêt dans les files d'attente avant l'embarquement dans le bateau. Il a été remis un questionnaire par véhicule de tourisme. Le questionnaire était rempli par le ménage dans le véhicule durant le temps d'attente. La récupération des questionnaires par les enquêteurs s'est fait en parcourant les files d'attente avant la montée dans le bateau. En ce qui concerne les passagers piétons, l'enquête a eu lieu dans la salle d'embarquement. La remise des questionnaires a eu lieu juste après la porte d'entrée dans le terminal. La récupération des questionnaires a eu lieu dans la salle d'embarquement, par un retour spontané de l'enquêté ou une récupération des questionnaires par l'enquêteur. La vacation enquêteur pour un bateau enquêté était 2 heures environ (arrivée 1h45 avant le départ prévu, et 15 mn de temps d'approche et de classement, comptage des questionnaires).
- Pour les aéroports : Il s'agissait d'enquêter tous les avions dont l'heure de départ prévue se situait dans une plage horaire de 2 heures. Les horaires de départ s'étalant de 6h00 à 24h00, il y avait donc 9 tranches horaires par jour. Deux ou trois enquêteurs, selon l'affluence, étaient nécessaires. L'enquête a eu lieu dans la salle d'embarquement. La remise des questionnaires a eu lieu juste après les portiques de sécurité. La récupération des questionnaires s'est faite par un retour spontané de l'enquêté ou par une récupération des questionnaires par l'enquêteur lors de l'embarquement dans l'avion. La vacation enquêteur pour une tranche horaire enquêtée était au maximum de 3h30 (arrivée 1h15 avant le premier départ prévu de la tranche horaire, départ 15 mn avant à la fin de la tranche et le temps d'approche et de classement, comptage des questionnaires estimé à 30 mn).

Le temps de remplissage du questionnaire était de l'ordre de 8 minutes pour les non-résidents et de 1 minute pour les résidents. Les enquêtés l'ont rempli dans les salles d'embarquement ou dans les files d'attente des véhicules.

2.4. Suivi de l'enquête : identification des trajets

Que ce soit pour le maritime ou l'aérien, l'identification des trajets enquêtés a été effectuée par les enquêteurs via la constitution de paquets de questionnaires pour chaque trajet enquêté. Ces paquets ont alors été identifiés, via une fiche suiveuse sur laquelle se trouve les identifiants de trajet à savoir :

- pour les bateaux : l'identifiant de la traversée, le nom de la compagnie, la date et l'heure prévue de départ et les ports de départ et de destination.
- pour les avions : l'identifiant du vol, le n° de vol, la date et l'heure prévue de départ de l'avion et les aéroports de départ et de destination.

Cela a nécessité, pour chaque plage de 2 heures enquêtée, de trier l'ensemble des questionnaires par vol. Ce tri a pu s'effectuer soit directement lors de la récupération des questionnaires, soit, ex post, en utilisant les réponses à la première question du questionnaire (destination, heure de départ prévue).

L'enquêteur a indiqué sur la fiche suiveuse le nombre de questionnaires correspondant. Ce décompte a été transmis quotidiennement à l'ATC et à l'Insee via une application spécifique dédiée à la gestion d'enquête. L'enquêteur a fait parvenir à l'Insee les questionnaires remplis via un envoi sécurisé de la poste.

L'acquisition des données des questionnaires a été réalisée par un prestataire, reconnu, de saisie optique.

L'Insee a archivé l'intégralité des questionnaires dans la salle d'archive de la direction régionale avant envoi par transporteur au prestataire de saisie optique. Les données sous forme de fichier ont été renvoyées par le prestataire via un espace sécurisé (SEF) où seul l'Insee pouvait avoir accès. Les questionnaires papiers ont été ensuite détruits par le prestataire.

2.5. Résultats de la collecte

72 000 questionnaires ont été recueillis parmi les 410 bateaux et 2200 avions enquêtés sur la période. Cela a été conforme aux estimations élaborées à partir des tests.

Les principales difficultés rencontrées durant la collecte sont de trois ordres :

- Tout d'abord le recrutement était de la compétence de l'ATC, l'Insee intervenait uniquement dans la formation, l'accompagnement et le suivi des enquêteurs afin de s'assurer de la qualité. Durant la saison, 13 enquêteurs ont été nécessaires sur le terrain. Recruter des personnes disponibles à des horaires atypiques (tôt le matin ou tard le soir) sur des jours non-réguliers durant la semaine et le week-end a été complexe. Par ailleurs, les conditions de travail ont été difficiles : les salles d'embarquement étaient surchargées et sur le port, la chaleur était importante. Ainsi, le turn-over était important. La réactivité INSEE-ATC a permis d'y faire face. Au total, 23 personnes ont travaillé sur l'enquête et 10 personnes ont abandonné en cours d'enquête. Des accompagnements INSEE-ATC ont été nombreux et réguliers.
- La seconde difficultés durant la collecte à laquelle nous avons dû faire face est le changement de planning des compagnies. En effet, les compagnies peuvent 14 jours avant la date prévue de la traversée modifier l'horaire prévu. En ce qui concerne l'aérien, la situation a été relativement simple puisque l'intégralité des vols d'une tranche horaires devaient être enquêtés. Ainsi certains vols non prévus ont pu être enquêtés. Le poids de sondage de la tranche horaire concernée leur a été affecté. En ce qui concerne le maritime, un bateau de Bastia a percuté un quai et n'a pas pu effectuer son programme de travail durant la saison. Les plannings ont été considérablement modifiés. Ainsi, il a fallu obtenir des compagnies qu'elles nous fassent parvenir les plannings réactualisés en fin de mois afin de demander à la DMCSI de retirer les échantillons sur le port de Bastia. Par ailleurs la collecte sur les ports s'est avérée très complexe lorsque le mode d'embarquement ne permettait pas de regrouper les voitures sur le quai. Ainsi les enquêteurs ont eu la possibilité de proposer un questionnaire simplifié composé de 10 questions importantes.
- Enfin, le suivi de l'enquête a été réalisé par l'Insee. Le suivi de la collecte, la permanence téléphonique, les accompagnements et formation des enquêteurs ont été chronophages et ont nécessité rigueur et disponibilité. Une permanence téléphonique a été mise en place afin de répondre aux difficultés des enquêteurs sur le terrain.

3. Méthodologie de l'enquête

3.1. Plan de sondage

3.1.1. En résumé

Les unités de sondage sont les bateaux d'une part, et les tranches horaires d'aéroport de deux heures d'autre part, afin de faciliter le protocole de collecte. Pour éviter un recouvrement entre ces deux échantillons, une pré-affectation des jours a été effectuée par un tirage équilibré à probabilités égales. Le tirage des tranches horaires et bateaux a ensuite été réalisé à l'aide d'un tirage stratifié, équilibré au sein de chacune des strates. Les allocations par strate ont été calculées par la DR de Corse en modifiant légèrement les allocations proportionnelles au nombre de passagers par strate déterminées par la division Sondages.

3.1.2. Stratification

Les deux bases de sondage ont été stratifiées par zone géographique croisée par le mois de l'année, car le nombre de départs et les caractéristiques des passagers variaient assez fortement selon la zone de départ et fortement selon le mois de l'année.

Les zones géographiques correspondent à un regroupement en 4 modalités des ports et des aéroports. Plus précisément :

1. La zone «Ajaccio» qui correspond au port d'Ajaccio pour la base des bateaux et à l'aéroport d'Ajaccio pour la base des avions ;
2. La zone «Bastia» qui correspond au port de Bastia pour la base des bateaux et à l'aéroport de Bastia pour la base des avions ;
3. La zone «Balagne» qui correspond au port de L'île Rousse pour la base des bateaux et à l'aéroport de Calvi pour la base des avions ;
4. La zone «Sud» qui correspond au port de Porto-Vecchio pour la base des bateaux et à l'aéroport de Figari pour la base des avions.

Ces zones géographiques ont ensuite été croisées avec le mois de l'année pour obtenir un total de 28 strates dans chaque base. Le port de Bonifacio (au Sud) étant plus particulier, il a constitué une cinquième zone de tirage à lui-seul.

3.1.3. Allocations

L'allocation de l'échantillon entre les strates a été calculée de façon à être proportionnelle au nombre de passagers par strates. Le total du nombre de nuits par avion\bateau étant fortement corrélé au nombre total de passagers présents dans l'avion, cette allocation contribue à une amélioration de la précision. Cependant des contraintes d'organisation de la collecte ont conduit à des modifications par rapport aux allocations optimales. Ces modifications étaient légères, et n'altéraient pas l'efficacité du choix de l'allocation.

3.1.4. Coordination négative des deux tirages

Pour éviter des collectes simultanées dans une même zone, étant donné que l'on ne dispose souvent que d'une équipe d'enquêteurs dans chaque zone, il est nécessaire que le recouvrement entre les deux échantillons soit faible voire inexistant. Une première solution a consisté à tirer ces deux échantillons de façon indépendante et à évaluer le taux de recouvrement. Lors des tests effectués, les taux de recouvrement journalier étaient de l'ordre de 40 % (dans une même zone géographique)

et les taux de recouvrement entre les demi-journées de l'ordre de 20 %, ce qui n'était pas acceptable en termes d'organisation de la collecte.

Il a donc été nécessaire de réaliser des tirages coordonnés négativement. Afin de réaliser cette coordination tout en conservant des bases équilibrées, nous avons réalisé une affectation aléatoire des jours x strates aux bateaux ou avions.

Afin d'éviter les recouvrements entre l'échantillon de tranches horaires d'avions et les départs des bateaux, nous avons donc tiré selon un tirage équilibré à probabilités égales les jours dédiés aux avions et les jours dédiés aux bateaux au sein de chaque strate. Les probabilités d'inclusion ont été fixés à 0,5 de façon à diviser la base des jours en deux partitions égales au sein de chaque strate. Les variables d'équilibrage mobilisées pour réaliser ce tirage au sein de chaque strate sont :

- le nombre de tranches horaires d'avion par jour ;
- le nombre de bateaux par jours ;
- le nombre de passagers transportés par avion par jours ;
- le nombre de passagers transportés par bateau par jour ;
- les sept indicatrices de jours de la semaine ;
- l'indicatrice de jour de vacances.

Il aurait été également possible de réaliser un tirage des jours avec stratification directe lors de la phase de vol de l'algorithme de tirage équilibré puis de réaliser la phase d'atterrissage de l'algorithme en regroupant l'ensemble des strates. Cela aurait facilité la phase d'atterrissage et aurait amélioré la qualité globale de l'équilibrage (meilleure répartition des jours au global) au prix d'une détérioration de cet équilibrage au sein des strates (estimations pour les passagers des bateaux et/ou des avions dans une strate spécifique de moindre qualité). Cette solution n'a pas été retenue, car les allocations par strate n'étaient pas nécessairement respectées dans ce cas.

Dans la zone comprenant Porto-Vecchio et Bonifacio, 4 journées par mois ont été sélectionnées selon un sondage aléatoire simple pour la collecte dans le port de Bonifacio. Ensuite, on tire aléatoirement le matin ou l'après-midi et on enquête l'ensemble des bateaux du matin ou de l'après-midi. Cette concentration de l'échantillon sur 4 demi-journées est motivée par le fait de faciliter la collecte dans ce port dont la capacité des bateaux est plus homogène et limitée.

3.1.5. Tirage des tranches horaires et des bateaux

À partir des jours tirés précédemment, on a construit une base restreinte des tranches horaires correspondant à l'ensemble des tranches horaires appartenant à un des jours tirés. Ensuite, au sein de chaque strate, les probabilités d'inclusion des tranches horaires ont été définies proportionnellement au nombre de passagers (en utilisant des taux de remplissage moyens)¹ car étant donné que le nombre de nuitées est fortement corrélé au nombre de passagers, cela contribue à une amélioration de la précision.

À partir de ces probabilités d'inclusion, on a réalisé **un tirage équilibré au sein de chaque strate** sur les variables suivantes :

- les probabilités d'inclusion (afin d'assurer le tirage d'un échantillon de taille fixe) ;
- le nombre de tranches horaires d'avion ;
- le nombre de passagers ;
- les indicatrices de tranches horaires regroupées ;
- les indicatrices de jours de la semaine.

¹ Pour une prochaine édition, les probabilités d'inclusions pourraient être calculées de façon proportionnelle au nombre de touristes dès lors que l'on disposera des informations plus précises sur les passagers (en distinguant les touristes, des voyageurs corses).

- l'indicatrice de jour de vacances
- le nombre de passagers transportés par une compagnie historique

Le poids de sondage associé à une tranche horaire tirée à l'issue des deux phases d'échantillonnage est égal à deux fois (poids de première phase) l'inverse de la probabilité d'inclusion de deuxième phase.

Enfin d'évaluer la qualité de ce plan d'échantillonnage, nous avons effectué 1 000 répliques de ce plan pour déterminer le CV et le biais associé à l'estimateur Horvitz-Thompson. On constate que le biais est quasiment nul, comme attendu. Les CV globaux sont compilés en Annexe 2 ; ils sont faibles, mais ils sont plus variables lorsque l'on regarde des résultats pour une strate (zone géographique X mois).

À partir des jours tirés précédemment, on a construit une base restreinte des bateaux correspondant à l'ensemble des départs de bateaux appartenant à un des jours tirés. Ensuite, au sein de chaque strate, les probabilités d'inclusion des bateaux ont été définies proportionnellement à la capacité des bateaux ou au nombre de passagers (en utilisant des taux de remplissage moyens).

À partir de ces probabilités d'inclusion, on a réalisé un tirage équilibré au sein de chaque strate sur les variables suivantes :

- les probabilités d'inclusion (afin d'assurer le tirage d'un échantillon de taille fixe)
- le nombre de départs de bateau
- le nombre de passagers
- les indicatrices de tranches horaires regroupées
- les sept indicatrices de jours de la semaine
- l'indicatrice de jour de vacances

Le poids de sondage associé à un bateau tiré à l'issue des deux phases d'échantillonnage est égal à deux fois (poids de première phase) l'inverse de la probabilité d'inclusion de deuxième phase de ce bateau.

Afin d'évaluer la qualité de ce plan d'échantillonnage, nous avons effectué 1 000 répliques de ce plan pour déterminer le CV et le biais associé à l'estimateur Horvitz-Thompson. On constate que le biais est quasiment nul, comme attendu. Les CV globaux sont compilés en Annexe 3 ; ils sont faibles, mais ils sont plus variables lorsque l'on regarde des résultats pour une strate (zone géographique X mois). En particulier, la variable destination, très corrélée à la zone géographique, ne peut donc pas être étudiée strate par strate.

3.1.6. Des améliorations du tirage pour la fin de la collecte

Deux problèmes se sont présentés en cours de collecte :

- le premier réside dans le fait que les jours de collecte n'étaient pas suffisamment espacés, ce qui pouvait générer une longue série de jours d'enquête pour les enquêteurs ;
- le deuxième problème réside dans le fait que deux bateaux ayant des heures de départ très proches pouvaient être sélectionnés, ce qui engendrait des difficultés de collecte.

Pour contourner ces deux problèmes, une méthode d'échantillonnage spatialement dispersée a été mobilisée. Cette méthode de tirage spatialement équilibré a été proposée par Grafström et Tillé (2013). L'idée générale de l'algorithme de tirage spatialement équilibré est de construire un cluster de $p+1$ (où p est le nombre de contraintes d'équilibrage) unités géographiquement proches, puis d'appliquer la phase de vol du cube sur ce cluster. Cela conduit à sélectionner au moins une unité

dans ce cluster en respectant les p contraintes localement dans ce cluster. Ensuite, les probabilités sont modifiées localement, ce qui assure que les probabilités d'inclusion des unités proches sont réduites, limitant ainsi la probabilité qu'une de ses unités proches soient sélectionnées par la suite dans l'étape suivante de l'algorithme.

Puis on répète la procédure : on sélectionne une unité, on crée un cluster de $p+1$ unités autour de l'unité sélectionnée et on applique la phase de vol du cube. On répète le processus tant qu'il reste plus de $p + 1$ unités. Pour finir, on applique la phase d'atterrissage classique du cube. La méthode de tirage spatialement équilibrée décrite ci-dessus est disponible dans le package R appelé *BalancedSampling* implémenté par Grafström (2016). Nous l'avons mobilisée dans notre contexte afin de disperser l'échantillon de jours non pas dans l'espace physique mais dans le temps. Ainsi, les séries de jours consécutifs étaient plus courtes et davantage espacées et cela nous a également permis de supprimer les sélections quasi-simultanées de bateaux.

3.2. Redressements

L'enquête a permis de collecter 72 000 questionnaires à partir desquels vont être estimés des résultats sur les clientèles touristiques et sur leurs pratiques (hébergement, etc.). Cependant, il reste à contrôler les données collectées, à prendre en compte les phénomènes de non-réponse qui ont pu impacter la collecte et à réaliser des calages ; pour cette dernière étape, le choix des marges de calage est très important, en particulier vis-à-vis des autres sources statistiques sur le tourisme.

Les opérations de redressements ont été effectuées en totalité au sein de la direction régionale de Corse, après une formation intensive aux traitements post-collecte réalisée en immersion à la division Sondages pendant quelques jours. Si les méthodes employées ne sont pas nécessairement novatrices, cette organisation montre qu'il est possible de réaliser un traitement de qualité de ces problématiques directement au sein de la direction régionale en charge de l'enquête.

3.2.1. Apurement

La première étape de traitement des questionnaires est l'apurement, c'est-à-dire la vérification de la cohérence des questionnaires en interne. L'idée ici est de corriger des erreurs de saisie optique ou de remplissage des questionnaires par les enquêtés dans les cas où l'erreur est manifeste ; par exemple, si le nombre de nuitées par type d'hébergement est au total plus important que le nombre de nuitées totales.

Il est également possible de mobiliser de l'information extérieure pour réaliser cet apurement. Par exemple, si l'on sait que les seuls vols pour lesquels la destination est l'aéroport de Marseille depuis la Corse sont ceux gérés par la compagnie Air Corsica, il est possible d'imputer cette information dès lors que l'on connaît la destination des passagers dans le questionnaire ; cette opération était utile principalement pour les vols ou traversées dont le planning était modifié.

Un dernier aspect traité au moment de l'apurement est celui de la détection des doublons de questionnaires. En effet, l'organisation de la collecte fait qu'un seul questionnaire doit être distribué et collecté pour chaque groupe de voyageurs, mais dans certains cas, plusieurs membres du groupe peuvent avoir complété le questionnaire par erreur. Il est important de détecter ces doublons car ils pourraient conduire à des erreurs d'estimation. Cette détection est assez simple à faire pour les grands groupes : on repère que deux questionnaires concernent un groupe de 17 personnes, par exemple, et en vérifiant qu'ils concordent on peut conclure qu'il s'agit probablement du même groupe et que l'on doit n'en conserver qu'un seul. Ce traitement est plus difficile à réaliser pour les

petits groupes, mais on peut faire l'hypothèse que dans ce cas, il doit y avoir moins de risques de doublons, car la communication en interne au groupe est plus simple.

3.2.2. Correction de la non-réponse

La question de la non-réponse totale ne se pose pas ici comme dans la majorité des enquêtes classiques ; en effet, notre échantillon est celui de bateaux et de tranches horaires au sein d'un aéroport, et il n'y a pas de non-réponse à ce niveau-là. La non-réponse totale peut provenir du refus ou de l'impossibilité de certains passagers à répondre à l'enquête, mais on ne dispose pas d'un listing des passagers, encore moins des groupes de passagers, qui auraient dû rentrer dans le cadre de l'enquête. Il n'est donc pas possible de réaliser de repondération au niveau groupe : cette non-réponse sera gérée à l'étape du calage.

Le traitement de la non-réponse consiste donc ici à gérer les problématiques de non-réponse partielle aux questions les plus importantes du questionnaire. Pour cela, la méthode de hot deck par classe est utilisée. Elle consiste à affecter à un questionnaire avec une réponse manquante une réponse donnée à un autre questionnaire qui partage les mêmes caractéristiques. L'imputation s'est fait de manière séquentielle : les variables sont d'abord regroupées par paquets, à l'intérieur duquel nous supposons ou savons qu'il existe de fortes corrélations, que des imputations séparées risqueraient de détruire ou fortement atténuer. Par exemple, les variables indiquant le nombre de nuitées par mode d'hébergement sont regroupées et imputées ensemble.

Ces paquets de variables sont classés par ordre d'importance décroissante, et chaque variable imputée à une étape peut servir à la construction des classes pour les variables suivantes.

Par exemple, pour la question du type d'hébergements, les classes ont été constituées en utilisant les variables suivantes : voyage professionnel ou non, nombre de personnes dans le groupe de voyageurs, destination du bateau ou de l'avion, compagnie, moyen de transport sur l'île, âge, durée depuis laquelle le séjour est réservé, etc...

La sélection des classes s'est faite à l'aide d'arbres de décision permettant d'identifier à chaque étape quelle était la variable permettant de mieux séparer les questionnaires de telle sorte à avoir des groupes homogènes en termes de taux de réponse et/ou de réponses. On aboutit à la fin à un ensemble de classes contenant environ une centaine d'individus, dans lesquelles on réalise un hot deck.

La méthode de hot deck utilisée ne permettait pas de choisir plusieurs fois le même questionnaire comme « donneur », c'est-à-dire comme valeur imputée pour un non-répondant. Elle consiste en effet à tirer dans chaque classe, contenant par exemple R répondants et N non-répondants, un échantillon de N donneurs parmi les R répondants par sondage aléatoire simple et d'affecter aléatoirement ces donneurs aux N non-répondants de la classe. Cette contrainte permet d'éviter de répliquer de nombreuses fois une réponse très atypique, de créer des résultats aberrants et de déformer les distributions des variables, mais elle peut entraîner des problèmes pratiques. En effet, il est possible que dans une classe le nombre de donneurs soit insuffisant pour réaliser l'imputation. Dans ce cas, les classes sont étendues en regroupant plusieurs classes de telle sorte à contenir 200 voire 400 individus par classe et un nombre de répondants supérieur au nombre de non-répondants dans chaque classe.

3.2.3. Calage

Le calage a été réalisé à l'aide de la macro CALMAR2 (Le Guennec, Sautory 2002). Plusieurs informations sont mobilisables ici : les informations issues de la base de sondage concernant le nombre de vols et de bateaux pour chaque destination à chaque période ; les informations issues des données des compagnies aériennes, et en particulier la répartition entre résidents corses et touristes pour les différents vols ; les données statistiques de tourisme connues à l'Insee via les autres dispositifs mis en œuvre tels que l'enquête de fréquentation des hôtels et des campings. Ce sont les deux premières sources qui ont été mobilisées pour le calage ; nous discuterons de la troisième dans un second temps.

La première approche utilisée était celle d'un calage au niveau groupe, c'est-à-dire de modifier les pondérations au niveau fin auquel est collectée l'information dans l'enquête de telle sorte à faire concorder la structure des répondants avec celle connue en termes de nombre de personnes par destination, par compagnie, par aéroport ou port de départ, par jour de la semaine, par mois, etc. Cette approche semble efficace, mais elle conduit à de très forts rapports de poids, c'est-à-dire des poids fortement modifiés, en particulier pour les groupes les plus grands. Ce résultat implique un risque de valeurs influentes pour ces groupes, qui risquent d'être atypiques. La solution n'est donc pas très robuste.

Il a donc été décidé de réaliser le calage au niveau trajet, c'est-à-dire vol ou bateau, ce qui permettait une dispersion des rapports de poids plus faible au prix d'une granularité moins fine en termes de prise en compte des caractéristiques des passagers. Or, une caractéristique au niveau groupe joue fortement sur la non-réponse à l'enquête et donc perturbe le calage au niveau trajet : il s'agit du fait d'être résident corse ou non. En effet, les résidents corses répondent très largement moins à cette enquête que les touristes. La repondération par calage réalisée au niveau vol conduit alors à augmenter le poids indépendamment du statut de résident corse ou non, et donc à augmenter artificiellement le poids des groupes de touristes, ce qui entraîne une surestimation du nombre de nuitées touristiques.

Pour pallier cet écueil, il est nécessaire d'intégrer des marges liées à cette caractéristique. L'information est difficile à obtenir, mais certaines des plus grandes compagnies aériennes ont communiqué à l'ATC (qui les a transmises à l'Insee) des données relatives à leurs passagers, et en particulier le caractère « résident » ou non de ceux-ci, en moyenne pour les vols du mois. Il y a ici une légère différence de concept. Dans les données fournies par les compagnies aériennes, on ne dispose pas du lieu de résidence des passages ; on va alors considérer qu'un passager est résident si son vol a été réservé avec un aller depuis la Corse et un retour vers la Corse (dans le cas inverse, on considérera qu'il s'agit d'un touriste), ou s'il a utilisé pour sa réservation le tarif « résident » (tarif négocié permettant un prix fixe pour les vols). Cette définition ne concorde pas exactement avec la notion utilisée dans le questionnaire (lieu de résidence) mais en semble un bon proxy. L'utilisation de ces marges dans le calage permet alors de réduire la surestimation évoquée au paragraphe précédent.

Une cinquantaine de marges ont donc été utilisées pour le calage. Il s'agit des principales destinations, compagnies, ports et infrastructures, ainsi que la part des résidents dans les plus grandes compagnies.

La question du calage sur d'autres sources statistiques se pose alors. On remarque en effet que les résultats mensuels de l'enquête auprès des voyageurs et de celle auprès des hôtels ou des campings ne concordent pas forcément ; on pourrait souhaiter caler les résultats de cet enquête sur l'autre pour en améliorer la robustesse. Cependant, plusieurs aspects peuvent expliquer ces écarts :

- Le champ temporel n'est pas exactement le même : un groupe quittant l'île au 1^{er} août sera comptabilisé dans les données d'août alors que la majorité de ses nuitées auront eu lieu en juillet ;
- Certains hébergements ne font pas l'objet d'enquêtes, par exemple les refuges et certains gîtes présents sur le circuit du GR20 ;
- Le concept de camping n'est pas nécessairement équivalent dans les deux enquêtes : par exemple, un camping « sauvage » consistant à garer son camping-car dans un endroit non spécifiquement prévu pourra être déclaré comme camping par le touriste, mais ne sera évidemment pas pris en compte par l'enquête auprès des campings officiels ;
- De même, certains concepts comme les hôtels peuvent différer entre les enquêtes, par exemple s'il s'agit de petits hôtels comportant peu de chambres qui peuvent être exclus du champ de l'enquête tourisme ;
- Il est possible qu'il y ait des biais de déclaration à la hausse ou à la baisse, que ce soit par les professionnels du tourisme ou par les voyageurs eux-mêmes.

Ainsi, ces différences entre les dispositifs font qu'il est intéressant d'étudier les différences entre les résultats qui seraient masquées artificiellement par un calage d'une source sur l'autre. Il a donc été décidé de ne pas intégrer ces marges dans le calage.

Bibliographie

- J. Le Guennec, O. Sautory : « Calmar2, une nouvelle version de la macro Calmar de redressement d'échantillon par calage », JMS 2002.
- Grafström, A. et Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.
- Grafström, A. Lisic, J (2016). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.2. <http://www.antongrafstrom.se/balancedsampling>

Annexe 1 – questionnaire de l'enquête

15 Comment avez-vous réservé votre hébergement ? (plusieurs réponses possibles)

- Pas de réservation Booking.com Air BnB, Aritel, Homelidays Le bon coin
 Directement auprès de l'hébergeur Agence de voyage, voyagistes Agence immobilière Autre

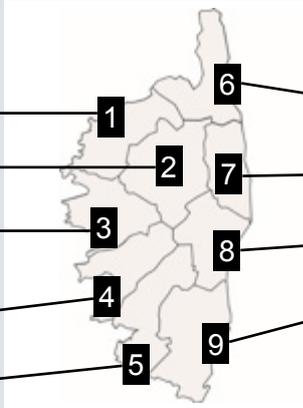
16 Au total, combien de nuits avez-vous passées en Corse lors de ce séjour ?

17 Comment se répartissent vos nuits dans ces différents modes d'hébergement (sur 2 caractères) ?

Mode d'hébergement	Nombre de nuits	Mode d'hébergement	Nombre de nuits
Chez la famille	<input type="text"/> <input type="text"/>	Chez des amis	<input type="text"/> <input type="text"/>
Dans votre résidence secondaire	<input type="text"/> <input type="text"/>	Location saisonnière (villa, appartement) auprès d'un particulier	<input type="text"/> <input type="text"/>
Hôtel	<input type="text"/> <input type="text"/>	Camping	<input type="text"/> <input type="text"/>
Résidence de tourisme	<input type="text"/> <input type="text"/>	Village de vacances	<input type="text"/> <input type="text"/>
Chambres et maisons d'hôtes et gîte	<input type="text"/> <input type="text"/>	Échange d'appartement, de maison	<input type="text"/> <input type="text"/>
Refuge	<input type="text"/> <input type="text"/>	Bateau de plaisance	<input type="text"/> <input type="text"/>

18 Combien de nuits avez-vous passées dans ces régions ?

Région	Nb. de nuits	Région	Nb. de nuits
1 Balagne, Calvi, L'île-Rousse, Galeria...	<input type="text"/> <input type="text"/>	6 Bastia, Saint-Florent, Cap-Corse, Patrimoine...	<input type="text"/> <input type="text"/>
2 Corte, Restonica, Niolu...	<input type="text"/> <input type="text"/>	7 Castagniccia, Casinca, Orezza, Costa Verde...	<input type="text"/> <input type="text"/>
3 Cargèse, Sagone, Porto, Piana...	<input type="text"/> <input type="text"/>	8 Côte Orientale, Ghisonaccia, Aléria...	<input type="text"/> <input type="text"/>
4 Ajaccio, Porticcio Pays ajaccien...	<input type="text"/> <input type="text"/>	9 Porto-Vecchio, Bonifacio Alta-Rocca...	<input type="text"/> <input type="text"/>
5 Sartène, Propriano, Taravo...	<input type="text"/> <input type="text"/>		



19 Comment vous êtes-vous déplacé durant votre séjour (plusieurs réponses possibles) ?

- Voiture personnelle Location de voiture Moto/Scooter Vélo
 Train Bateau de plaisance Camping-Car Véhicule aménagé
 À pied Transports en commun Taxi Car

20 Avez-vous pratiqué une des activités payantes suivantes ?

- Activités de pleine nature (accrobranche, canyoning, randonnée avec guide...) Nautisme, plongée
 Visite patrimoine naturel (excursion bateau, train...) Musée, exposition
 Visite patrimoine culturel/traditionnel (terroir, caves, artisanat) Concert

21 Quel est votre niveau de satisfaction selon les items suivants ?

	Rapport qualité/prix du séjour	Coût du transport	Coût de l'hébergement	Coût de la restauration	Qualité de l'accueil	Qualité hébergement	Qualité restauration
Très satisfait	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Plutôt satisfait	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Peu satisfait	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pas du tout satisfait	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Non concerné	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

22 Avez-vous l'intention de revenir en 2018 ?

- Oui, certainement Oui, probablement Je ne sais pas
 Non, certainement pas Non, probablement pas

23 Quels sont les revenus mensuels, en moyenne, du foyer (en euros) ?

- Pas de revenus Moins de 1500 euros 1501-2500 euros
 2501-4000 euros 4001-6000 euros Plus de 6000 euros

Annexe 2 – Coefficients de variation obtenus pour les vols

Strate	Mois	Zone	Capacité	Taille	HP	NHP	Destinations (regroupées en zones)				
							PACA	Paris	Étranger	Nord France	Sud France
		Corse – période complète	1,45	2,03	1,41	1,87	3,76	2,61	4,99	5,24	5,18
11	avril	Ajaccio	5,59	7,88	5,19	8,79	16,14	12,88	50,92	20,85	23,62
12	avril	Bastia	7,15	7,79	5,69	13,64	15,09	15,24	42,16	31,14	29,37
13	avril	Balagne	11,40	13,88	10,33	43,96	19,71	16,08	51,64	130,52	
14	avril	Sud	6,61	16,25	9,37	10,46	33,32	18,20	49,91	49,21	84,06
21	mai	Ajaccio	5,34	6,61	4,98	7,78	16,54	10,67	27,31	16,63	22,48
22	mai	Bastia	6,43	13,45	6,27	10,21	18,35	13,81	27,54	36,82	23,41
23	mai	Balagne	10,80	20,08	10,35	23,41	27,19	15,15	26,50	65,86	
24	mai	Sud	6,74	10,50	8,82	7,79	17,70	13,56	23,38	31,13	48,43
31	juin	Ajaccio	4,20	6,71	3,41	6,33	13,98	8,88	25,37	16,56	20,81
32	juin	Bastia	6,07	9,53	3,73	10,98	19,33	13,50	23,80	25,97	28,56
33	juin	Balagne	8,92	12,22	7,49	15,74	24,13	15,53	16,87	68,74	205,19
34	juin	Sud	7,06	8,25	6,00	10,68	14,63	11,49	21,45	23,82	32,57
41	juillet	Ajaccio	5,87	6,69	5,90	6,61	18,66	10,85	22,55	16,73	19,75
42	juillet	Bastia	5,90	8,66	6,45	7,44	21,83	11,81	13,85	26,50	22,66
43	juillet	Balagne	7,40	18,10	6,44	13,97	29,47	14,50	22,33	53,44	89,70
44	juillet	Sud	6,39	9,27	6,22	8,11	15,94	13,37	24,79	27,38	27,51
51	août	Ajaccio	4,44	6,03	5,69	3,91	15,40	9,38	22,35	16,99	17,26
52	août	Bastia	6,12	9,49	5,26	9,01	20,32	11,86	19,69	25,69	20,75
53	août	Balagne	7,82	10,89	6,87	15,98	29,45	12,54	21,50	51,55	63,40
54	août	Sud	5,28	11,36	5,85	6,71	15,80	10,54	30,35	24,46	22,50
61	septembre	Ajaccio	3,57	6,07	3,53	5,38	14,63	9,73	26,79	15,03	16,76
62	septembre	Bastia	4,84	10,69	3,18	8,61	20,34	13,10	20,33	20,26	24,97
63	septembre	Balagne	7,36	9,46	6,46	16,09	24,86	13,98	21,92	64,58	191,72
64	septembre	Sud	5,35	8,13	5,44	7,44	15,32	11,77	20,77	21,95	22,95
71	octobre	Ajaccio	5,50	8,19	5,73	8,02	16,10	11,74	26,38	21,97	24,81
72	octobre	Bastia	7,23	8,88	5,39	13,02	16,28	18,20	27,76	45,55	28,79
73	octobre	Balagne	10,05	13,51	7,31	24,84	20,17	12,04	23,13	107,30	
74	octobre	Sud	9,52	13,69	10,36	14,35	34,18	22,27	35,15	57,72	55,45

Annexe 3 - Coefficients de variation obtenus pour les traversées

Strate	Mois	Zone	Capacité	Taille	Destinations (regroupées en zones)			
					Corse	France	Italie	Sardaigne
		Corse – période complète	1,64	1,97	76,96	4,58	5,72	3,38
11	avril	Ajaccio	3,50	8,67		8,67		
12	avril	Bastia	2,95	6,88		15,86	19,49	220,83
13	avril	Balagne	15,70	24,47	95,84	28,20	164,07	
14	avril	Sud	7,85	7,53		31,87		8,12
21	mai	Ajaccio	5,42	7,59		7,59		
22	mai	Bastia	5,07	7,89	245,10	22,16	16,76	
23	mai	Balagne	26,38	21,72	123,54	21,53		
24	mai	Sud	10,84	7,01		23,45		7,21
31	juin	Ajaccio	3,16	10,98		13,61	173,83	
32	juin	Bastia	2,87	5,81		22,60	15,92	322,38
33	juin	Balagne	16,60	32,70	191,81	33,20		
34	juin	Sud	8,89	4,84		30,20	156,90	5,73
41	juillet	Ajaccio	4,68	12,19		12,19		
42	juillet	Bastia	4,00	4,67		15,73	11,62	
43	juillet	Balagne	7,45	22,78		25,23	192,46	
44	juillet	Sud	5,30	6,63		38,63		7,53
51	août	Ajaccio	3,95	11,76		11,76		
52	août	Bastia	4,46	5,07		22,31	11,56	
53	août	Balagne	8,53	20,68		26,61	135,63	
54	août	Sud	4,96	6,08		36,36	203,57	7,52
61	septembre	Ajaccio	3,59	8,72		8,72		
62	septembre	Bastia	3,47	7,20		31,50	14,91	136,84
63	septembre	Balagne	12,20	27,02		30,21	143,62	
64	septembre	Sud	3,12	4,94	327,81	96,82		8,59
71	octobre	Ajaccio	4,93	7,00	195,89	8,61		217,98
72	octobre	Bastia	5,54	8,53	229,75	18,53	17,59	
73	octobre	Balagne	20,21	16,27		18,05	103,28	
74	octobre	Sud	5,54	10,88	255,33	66,34	664,38	13,47