GÉOSTATISTIQUE

La pratique de la Géostatistique est, d'abord, une occasion de rencontres: rencontres entre des champs d'application variés et parfois inattendus, entre des problématiques sans cesse renouvelées, et également une confrontation entre des objectifs et contraintes purement techniques d'une part et les exigences sociales, économiques, environnementales d'un monde complexe d'autre part. Autrement dit, tout en étant fiers de ce que le néologisme "géostatistique" jadis forgé à l'École des Mines ait trouvé droit de cité dans le Petit Larousse, il est satisfaisant d'observer, au fil des ans, que l'immuable définition qu'en donne le dictionnaire s'éloigne de plus en plus de la réalité, et que notre discipline trouve à s'exprimer bien au-delà de la simple estimation des gisements miniers. De fait, dans tout domaine où des jeux de données numériques présentent une organisation spatiale ou temporelle, la Géostatistique a les outils pour apporter un éclairage original, à la fois constructif et sans concession. Il semble que cet aspect transversal et non-conformiste de la Géostatistique constitue désormais son caractère dominant au regard des optionnaires, et nous ne manquerons pas dans le futur de justifier cette appréciation. Ainsi, chaque année, la diversité des vœux des étudiants constitue une chance exceptionnelle de tester des méthodes nouvelles et de parcourir des domaines nouveaux, et la garantie d'insuffler un surplus de dynamisme à l'équipe encadrante. Mais la mise en œuvre d'une Géostatistique de qualité exige en permanence d'assurer un équilibre, parfois délicat, entre des exigences souvent contradictoires: garantir une rigueur théorique indispensable à la fiabilité des résultats tout en conservant un point de vue pragmatique et réaliste afin que les conclusions abstraites trouvent à s'appliquer sur le terrain. Sans oublier une indispensable déontologie, dans des domaines où souvent les contraintes économiques ou environnementales soumettent le géostatisticien à des pressions qui ne relèvent plus de la science ou de la technique...

Ouverture et équilibre: c'est dans cet esprit que nous avons continué à proposer à la promotion actuelle un voyage de deux semaines en Guyane où, dans le contexte inhabituel et parfois tourmenté d'un DOM, les visites à des laboratoires, à des industriels et à des organismes institutionnels ont permis tout à la fois d'élargir l'horizon des optionnaires et de susciter un échange ouvert et fructueux avec nos interlocuteurs. Partie intégrante de la scolarité, la mission en Guyane constitue pour les optionnaires la phase d'initiation à la réalité du terrain.

Enfin, le souci d'ouverture s'est à nouveau exprimé au niveau des sujets de l'option. Outre la variété des champs d'application, une multiplicité de méthodes statistiques est employée. Nous avons ici l'illustration que notre démarche méthodologique peut trouver à s'appliquer dans de multiples domaines et cela souligne le caractère généraliste de l'option Géostatistique et Probabilités Appliquées, tant en ce qui concerne les champs applications abordés que les méthodes mathématiques mises en jeu.

Hans WACKERNAGEL

GEOSTATISTICS

The aim of Geostatistics is to study quantitative phenomena that are structured in space and/or time. Engineers are almost inevitably faced with problems related to geostatistical techniques such as the evaluation of natural reserves, the analysis of time series, cartography, etc., and, broadly speaking, any processing of "regionalized variables" according to the terminology of G. Matheron, the founder of geostatistics.

The Geostatistics and Applied Probability Minor gives priority to probabilistic models and statistical methods, and in particular focusses on their application to the processing of spatial and temporal data.

The training in this Minor is aimed essentially at providing students with a critical mastering of some of the tools which they might need to use during their internship. As an introduction to "geostatistics in the field", the Minor provides an opportunity of entering into contact with companies and personnel working with geostatistics in fields of application that correspond as much as possible to the themes of particular interest to the students. It is essential that the students acquire a sense of balance between an empirical approach and a mathematical approach to a problem, a sense of balance which the third-year internship will put into practice in real terms.

The presentations this year again reflect the great variety of themes and applications that are encountered in this branch of applied statistics.

Hans WACKERNAGEL

GÉOSTATISTIQUE *GEOSTATISTICS*

Mardi 28 juin *Tuesday 28th June*

V106B

Mike PEREIRA

11h-12h

Estimation statistique du risque routier

Statistical estimation of road risk

Les entreprises possédant des flottes de véhicules motorisés importantes, ainsi que les acteurs de l'aménagement territorial ou même les assureurs s'intéressent à la notion de risque routier. Le risque routier est défini comme le risque d'être victime d'un accident de la circulation lors d'un déplacement sur la voie publique. Il est estimé par le rapport, dans une zone donnée, de la densité d'accidents s'y étant produits sur une mesure d'exposition au risque. Ce taux permet de normaliser l'information du nombre d'accidents se produisant dans une zone, par un facteur tenant compte de sa propension à produire des opportunités d'accidents.
Deux problèmes d'estimation statistique se posent alors, tous deux contraints au réseau routier (les variables évoluent le « long des routes »). Pour la densité d'accidents, trois méthodes d'estimation sont comparées : le krigeage du nombre d'accidents, l'estimation par noyau et l'inférence bayésienne par simulations MCMC. Pour la mesure d'exposition au risque, on se concentre sur l'estimation d'un facteur déterminant : le trafic routier. N'étant pas disponible de manière exhaustive sur le réseau, on en compare l'estimation par krigeage et par simulations MCMC (conjointes à celles de la densité). L'objectif du stage est de produire une méthode efficace d'estimation du risque routier sur une portion de réseau donnée, permettant la création de cartes de risque.

Companies with a large amount of motorized vehicles, organisms in charge of land-use planning and also insurance companies are interested in the notion of road risk. Road risk is defined as the risk of getting involved in a car crash while travelling on public roads. It is estimated by the ratio, in a given area, of the density of crashes that occurred there by a measure of risk exposition. This ratio allows a normalization of the number of crashes occurring in an area by a factor that takes into account its tendency to produce crash opportunities.
Two problems of statistical estimation arise, but both are constrained to the road network (given that the estimated variables vary "along the roads"). Regarding the crash density, three methods of estimation are compared: kriging of the number of crashes, kernel estimation and Bayesian inference using MCMC simulations. Regarding the risk exposition evaluation, the focus is centered on the estimation of a decisive factor: road traffic. Given that it is not available exhaustively on the network, its estimation by kriging and by MCMC simulations (coupled to those of the crash density) are compared. The aim of the internship is to produce an efficient method of road risk estimation on a given portion of the road network, allowing the creation of risk maps.

Estimages
Paris

Julien BARRAU

13h30-14h30

Data mining: prédiction des revenus d'un jeu vidéo et classification des utilisateurs

Data mining: lifetime value prediction and video game user classification

Avec l'essor des objets connectés et le développement des technologies big data, un nombre toujours croissant de données informatiques sont recueillies, mais bien souvent, seule une fraction d'entre elles sont exploitées, alors qu'elles contiennent des renseignements potentiellement précieux concernant les utilisateurs d'un site web, d'un service, ou en l'occurrence d'un jeu vidéo.

With the advent of the internet of things and the development of big data technologies, an ever-increasing amount of data is collected, but often only a small part of it is actually used, even when they may reveal some valuable information about the users of a web-page, a service, or in this case a video game.
How much income (advertising income and tran-

Quels seront les revenus (publicitaires et transactions) générés par l'ensemble des personnes qui ont installé l'application au mois de juin ? Sachant que sur la même période, X dollars ont été dépensés en marketing (X généralement grand), cet investissement sera-t-il rentabilisé ?
Certains pays seront-ils plus rentables que d'autres ? À partir des données récoltées durant les premiers jours, comment prévoir la durée (dite de rétention) pendant laquelle un joueur restera actif ? Quels facteurs peuvent influencer la rétention de ce joueur : son pourcentage de victoires ? Le temps de réponse de l'adversaire ? Le jour d'installation de l'application ? L'âge de son chien ? Toutes ces questions ont une importance stratégique pour orienter les investissements de l'entreprise et guider le développement du jeu. C'est pour y répondre que les algorithmes de machine learning les plus récents ainsi que diverses méthodes statistiques ont été étudiés et mis en œuvre.

sactions) will be generated by all the people who installed the application in a given period? Provided that X dollars were spent on marketing over the same period (X typically large), is this investment going to pay off? Are some countries going to be more profitable than others? Given the data collected during the first days, how can we predict how long a player will remain engaged (retention time)? Which factors can impact the retention time of this player: his winning percentage? His opponent's response time? The day when he installed the application? His favourite tea brand? All these issues are of strategic significance to adjust the company's investments and to guide the game's development. In order to address them, the latest machine learning algorithms as well as various statistical methods were studied and implemented.

IsCool Entertainment
Paris

14h30-15h30

Création de critères marketing et d'estimateurs statistiques associés pour l'optimisation de campagnes publicitaires

Creation of marketing profile estimators for the optimization of mobile advertising campaigns

Alexandre BENCHAOUINE

Databerries est une entreprise qui collecte des données de géolocalisation pour ensuite adresser de la publicité ciblée. La valeur ajoutée de l'entreprise repose sur le fait qu'elle peut effectuer une mesure des visites dans les magasins et ainsi déterminer le nombre de visites incrémentales dues à la publicité. Le business model de Databerries est donc de collecter des données de géolocalisation, créer des estimateurs de profils marketing intéressants et de vendre de la publicité ciblée vers ces profils. Dans un second temps, Databerries se rémunère en partie sur le nombre de visite incrémentales générées dans les magasins. Il est donc important de maîtriser tous les estimateurs de la chaîne de création de valeur pour avoir une facturation exacte. L'objectif de mon stage a été d'en comprendre les travers et de chercher à les optimiser en comprenant d'où venait le biais intrinsèque à tout estimateur.

Databerries is a company that collects geolocation data in order to target users with relevant advertisement. The company's added value relies on the possibility to measure the visits in stores and then to identify the incremental visits due to the targeted advertisement. Databerries business model is to collect geolocation data and create interesting marketing profile estimators to sell targeted advertisement. Furthermore, Databerries is partially making money thanks to the incremental visits in stores generated with the advertisement. Thus, it is important to master all the estimators that Databerries uses in the value chain to produce exact billings. The aim of my internship is to understand and optimize these estimators.

Databerries
Paris

15h30-16h30

Analyse de données GPS pour un outil publicitaire

GPS data analysis for an advertising software tool

Oscar ROCHE

L'analyse d'ensembles importants de données met en jeu des techniques novatrices à la fois au niveau hardware qu'au niveau software. Chez Databerries, des ensembles de données contenant des localisations GPS d'utilisateurs de smartphones sont traités à l'aide de requêtes SQL sur la plateforme Google BigQuery. Il s'agit d'effectuer des statistiques sur ces données, afin

Data analysis on large datasets requires innovative techniques, both from hardware and software standpoints. At Databerries, datasets containing GPS locations belonging to millions of smartphone users are queried using the SQL language on the Google BigQuery database system. These users are to be targeted using statistics on these datasets, in order to measure their likelihood of

de cibler des utilisateurs potentiellements intéressés par des campagnes publicitaires, puis de croiser ces données avec un suivi afin de mesurer l'efficacité des campagnes.
Les modèles mathématiques utilisés pour le ciblage autant que pour le suivi comportent des paramètres définis de façon empirique : des analyses de sensibilité sur les résultats permettent de recalibrer ces paramètres. Un autre aspect important et parfois délicat de l'analyse de données est la mise en forme des données : même si ces dernière sont disponibles, elles ne le sont souvent pas dans des formats compatibles ou de façon directement intelligible. Un exemple de solution à ce problème est la programmation de scrappers visant à récupérer des adresses d'enseignes sur des sites Internet en mettant en oeuvre, outre des compétences de programmation, des concepts théoriques ayant trait aux expressions régulières et, plus largement, au pattern matching afin que les données soient utilisables dans les bases de données SQL.

interest in some given advertising content and to subsequently measure the efficiency of the advertising campaigns. The mathematical models used for targeting users as well as following the adverstising campaigns are defined empirically : sensitivity analyses on the results allow to recalibrate these parameters.
Another major, and somewhat tricky aspect of data analysis is data formatting : even when data is available, it is very seldom happens to be in a suitable format, in the worst case it may even not be intelligible. An example of a solution to this problem is the programming of website scrappers in order to gather store adresses on internet websites ; this uses theoretical concepts related to regular expressions - more generally pattern matching in addition to programming skills for the data to be usable in SQL databases.

Databerries
Paris


Charbel
EL HACHEM

16h30-17h30

Optimisation du contenu d'un site web de média et segmentation de son audience

Content optimization and visitor segmentation of a news website

L'analyse des données devient aujourd'hui un facteur de succès important pour les médias en ligne. Ces médias, notamment Slate cherchent principalement à créer du contenu qui soit lu par le plus de gens possible. Le but de ce stage est de fournir des informations qui pourraient aider l'équipe éditoriale dans le choix du contenu : quels articles publier, quelles vidéos ? Il s'agit donc en partie de comprendre les préférences de l'audience du site en matière de sujets et de tons. Il est donc utile de définir des variables pour ces articles ,afin de les caractériser au mieux. Pour cela il existe dans la base de données des tags pour chaque article, ces tags sont souvent des expressions de deux à trois mots expliquant le thème de l'article. D'autres variables sont envisagées comme les heures auxquelles sont publiés les articles au quotidien, la taille des articles etc. D'autre part, le succès des articles est mesuré par des solutions de Web Analytics, qui fournissent des données importantes : le temps passé par page, les niveaux de clics, le niveau d'engagement, etc.
Le deuxième objectif est d'essayer de comprendre l'audience, de caractériser les utilisateurs en matière d'âge, de sexe, de niveau de salaire, de lieu d'habitation, ces données étant récoltées grâce aux cookies. Cette dernière analyse est utile pour pouvoir bien commercialiser le site auprès des marques qui cherchent à faire du marketing en ligne.

Data analysis has become a decisive factor when it comes to understand the success of news websites. News websites such Slate.fr seek primarily to create content that attracts a lot of web surfers. The purpose of this internship is to provide information that could help the editorial team in the choice of content. Hence trying to answer questions like: what type of articles and videos do people enjoy on the website? It is therefore useful to define variables for these items in order to characterize best as possible. The database of the website is very useful for this, it has tags for each article. A tag is a word or an expression that gives information on the main content of the article and on the subject that is discussed. Other variables like the size of the article and the time ofpublishing are also taken into consideration. On the other hand the success of the articles is measured by web analytic solutions that provide important data on visitor's behavior: time spent per page, the levels of clicks, the level of engagement etc.
The second objective is to try to characterize the audience in terms of age, sex, salary, and place of residence using data collected through cookies. The latter analysis will be useful for marketing and business purposes.

Slate.fr
Paris

9h30-10h30

Segmentation radar Doppler de fouillis de mer par méthode mean-shift

Doppler spectrum segmentation of radar sea-clutter by mean-shift

Thibault FORGET

Les inhomogénéités longitudinales de fouillis de mer Radar sont caractérisées par des variations en moyenne et en largeur du spectre Doppler. L'objectif du stage est de proposer une nouvelle approche pour l'estimation robuste de densité et la segmentation des spectres Doppler du fouillis de mer. Pour chaque case distance, le signal Doppler est caractérisé par une matrice hermitienne Toeplitz définie positive représentée dans le polydisque unitaire de Poincaré et une adaptation des méthodes à noyaux est utilisée pour l'estimation de densité sur cette variété Riemanienne particulière. A partir de cette approche non-paramétrique d'estimation de densité des spectres Doppler, la segmentation des données s'effectue par extension des méthodes de classification « mean-shift » sur ces densités sur le polydisque unitaire de Poincaré. Cette segmentation statistique est nécessaire pour une détection de cible robuste en fouillis de mer, en particulier en cas de mer agitée.

Radar sea-clutter inhomogeneity in range is characterized by Doppler mean and spectrum width variations. The aim of the internship is to propose a new approach for robust statistical density estimation and segmentation of sea-clutter Doppler spectrum. In each range cell, Doppler is characterized by a Toeplitz Hermitian positive-definite covariance matrix that is coded in Poincaré's unit polydisk and we use adaptation of standard kernel methods to perform density estimation on this specific Riemannian manifold. Based on this non-parametric approach to estimate statistical density of Doppler Spectrum, we address the problem of sea-clutter data mapping and segmentation by extending the "mean-shift" tool for these densities on Poincaré's unit polydisk. This statistical segmentation is requested for robust detection of targets in sea-clutter, especially in the case of high sea state.

Thalès Air Systems
Limours

10h30-11h30

Modèle de la fibrillation auriculaire basé sur la population : analyse statistique de la forme et physiologie

Population-based model of atrial fibrillation: from shape statistics to groupwise physiology

Shuman JIA

La fibrillation auriculaire (FA) est l'arythmie cardiaque la plus courante, caractérisée par une activation électrique chaotique des oreillettes, empêchant une contraction synchronisée. Cette maladie touche actuellement plus de 6 millions d'Européens. Ses complications potentiellement mortelles et sa progression rapide appellent un diagnostic aussi tôt que possible et un traitement efficace. L'ablation chirurgicale, une procédure invasive qui établit des lésions sur des zones déterminantes dans le déclenchement de la FA et crée une barrière à la propagation de l'arythmie, est un traitement efficace. Cependant, l'efficacité de la première ablation peut varier de 30% - 75%, de telle sorte que plusieurs procédures d'ablation peuvent être recommandées et la fonction mécanique auriculaire peut être lésée. Actuellement, il n'y a pas d'outil qui permette aux cliniciens d'accéder aux données intégrées du patient atteint de la FA ainsi que des modèles prédictifs, afin de faciliter la stratification du risque et la planification du traitement ultérieur. Afin de comprendre quels patients sont de bons

Atrial fibrillation (AF) is the most common cardiac arrhythmia, characterized by chaotic electrical activation of the atrial, preventing synchronized contraction. This epidemic currently affects more than 6 million Europeans and its life-threatening complications and fast progression call for as early as possible diagnosis and effective treatment. Atrial ablation, an invasive procedure which establishes lesions to block the trigger points of AF and creates a barrier to the propagation of the arrhythmia, is an effective treatment of AF. However, the efficacy of first time ablation may range from 30% - 75%, such that multiple ablation procedures may be recommended and atrial mechanical function may be adversely affected. Currently there is no decision support system available, enabling clinicians to access integrated AF patient data together with predictive models to facilitate risk stratification and subsequent treatment planning. In order to understand which patients are good candidates for atrial ablation and which patients are at risk for arrhythmia recurrence, we analyze, in collaboration with cli-

candidats pour l'ablation auriculaire et quels patients sont à risque d'arythmie récidive, nous analysons, en collaboration avec des cliniciens, une base de données de plus de 150 images MDCT 3D et utilisons des outils statistiques de pointe pour étudier les corrélations entre la forme des oreillettes et des facteurs cliniques, y compris l'âge, le sexe, le type d'arythmie, la durée de FA, la survenue d'accident vasculaire cérébral, les récidives après l'ablation, etc. Finalement, cette analyse pourrait être étendue à l'analyse statistique de la déformation des oreillettes entre des sous-groupes de patients et aux modèles auriculaires paramétriques.

nicians, a database of 150+ MDCT 3D images and use state-of-the-art shape statistics tools to explore correlations between atrial shape features in this patient population and clinical factors including age, gender, type of arrhythmia, AF duration, occurrence of stroke, post-ablation recurrences… Eventually, this analysis could be extended to deformation-based statistical analysis among the patient groups and parametric atrial models.

Inria
Sophia-Antipolis

Adrien
DE LA VAISSIÈRE

11h30-12h30

Témoins virtuels pour l'étude de facteurs environnementaux de maladies

Virtual controls to study environmental parameters of a disease

Pour repérer les facteurs environnementaux d'une maladie, il existe plusieurs méthodes épidémiologiques classiques. La plus courante est l'étude cas-témoins. On dispose d'un groupe d'individus malades, les cas, dont on a les adresses. A chaque adresse et donc à chaque individu, nous allons pouvoir associer des données environnementales. Puis à chaque cas nous allons apparier un individu non malade, un témoin, dont on connaît aussi l'adresse et donc les données environnementales. Finalement nous allons réaliser des études statistiques avec les données de ces deux groupes, afin de faire ressortir ou non certains facteurs environnementaux propices à la contraction de la maladie.
Lorsque des résultats de ce genre d'études paraissent, on accuse le choix des témoins: "Et si on les avait choisis différemment?". On parle de biais de sélection. Pour essayer d'éviter ce biais, nous travaillons sur un nouveau concept: les témoins virtuels. A la place d'utiliser des témoins physiques avec toutes les contraintes que cela entraîne, nous allons écrire un algorithme qui va associer à chaque cas, un "témoin virtuel" qui sera soumis à un certain environnement à partir d'un emplacement géographique tiré suivant certaines conditions. Cela permettrait de tirer de nombreux jeux de témoins, et d'avoir plus de résultats. Mais des questions persistent. Quel algorithme choisir pour trouver un bon témoin (qui, s'il avait été malade aurait été un cas)? Suivant quels critères? Quelles proportions? Quand est-ce que l'on considère qu'un algorithme est bon ? Ce sont les questions auxquelles je vais essayer de répondre en appliquant ces méthodes à des exemples concrets.

There are a few classical methods to spot environmental factors of a disease. The most frequently used is the case-control study. We have one group of people who have the disease, the cases, with their addresses. To each address, and thus to all cases, we will associate environmental data. Then, to each case, we will associate a healthy individual, a control, of which we know the address and thus the environmental parameters. Finally, we are going to carry out statistical analyses with the data of these two groups, in order to underline whether certain environmental parameters are likely to cause the contraction of the disease. When this kind of results is published, the choice of the controls is always blamed: "what if it they had been chosen differently?". This is referred to as selection bias. To try to avoid this bias, we are working on a new concept: virtual controls. Instead of using real, human controls with all the limitations that it involves, we are going to write an algorithm which will associate, to each case, a virtual control and an environment thanks to a location, drawn at random under certain conditions. This would provide us with a larger number of controls and more results. Nonetheless, some questions still remain. Which algorithm should be chosen to find a good control? What is a good virtual control? Under which constraints? These are issues that I will address by applying these methods to specific examples.

Inserm
Kremlin-Bicêtre

Adaptation de méthodes géostatistiques pour grands jeux de données

Adapting geostatistical methods for big datasets

Jean-Baptiste CHEVREL

L'utilisation d'estimateurs performants en exploration minière est une problématique centrale, particulièrement dans ce secteur exposé à une conjoncture économique difficile. Les données de teneur collectées en certains endroits par les grandes compagnies minières peuvent être utilisées pour estimer ces teneurs en tout point du champ. Il en résulte bien souvent de grands jeux de données qui rendent impossible l'utilisation des méthodes géostatistiques classiques (krigeage).
Il s'agit donc de trouver des méthodes mathématiques permettant d'estimer des interfaces entre domaines géologiques, à partir de grands jeux de données, et dans un temps acceptable. Les mesures de teneur collectées par les grandes sociétés minières sont bien souvent chères et difficiles à réaliser, ce qui rend inacceptable le fait de ne pas prendre toutes ces données en compte. La méthode retenue devra attribuer un rôle privilégié à certains points de données, sans pour autant exclure de donnée.

The use of efficient estimators in mining exploration is a prime concern, especially in the challenging economic environment faced by this sector. The grade data measured at a number of locations by the large mining companies can be used to estimate the grade at any point of the field. These datasets are often massive, due to the size of the fields of interest. These massive datasets make it often impossible to us classical geostatistical methods such as kriging.
The idea is therefore to find alternative mathematical methods enabling the estimation of interfaces between different geological domains using big datasets. This has to be done in a relatively limited time to make a method efficient from the user's standpoint. The grade data are often difficult and expensive to collect for the mining companies. Therefore it is unacceptable to eliminate data. The method retained will consequently have to give certain points a privileged role without excluding completely any data from the estimation.

Géovariances
Avon