GÉOSTATISTIQUE *GEOSTATISTICS*

Mardi 27 juin *Tuesday 27th June*

L213

Soufiane MOURRAGUI

8h30-9h30

Projections aléatoires pour la classification de données génomiques

Statistical estimation of road risk

Les progrès techniques des dernières années en matière de séquençage d'ADN ont donné accès à de très abondantes données. Une des problématiques de l'analyse des données issues de ces méthodes de séquençage à haut débit est le nombre important d'échantillons récoltés et la difficulté de les classer efficacement et rapidement. Ce problème se pose notamment en métagénomique, où des millions de séquences courtes doivent être attribuées aux génomes des bactéries dont elles proviennent. Les méthodes usuelles, comme les SVM, reposant sur l'astuce du noyau posent des problèmes de calcul et de stockage importants ; il en va de même pour les méthodes reposant sur la comparaison directe entre les différentes séquences.
Les méthodes utilisant l'astuce du noyau reposent sur le fait que tout noyau peut être écrit comme un produit scalaire dans un espace de Hilbert. Cet espace pouvant être de très grande dimension, souvent même infinie, l'astuce du noyau repose sur une utilisation implicite de cet espace sans calcul direct des vecteurs. Une approche nouvelle proposée il y a une dizaine d'années est de rendre la complexité linéaire par randomisation. L'espace dans lequel le noyau est représenté est alors approché par un espace de dimension finie dans lequel des techniques rapides d'optimisation comme la descente de gradient stochastique peuvent être utilisées. De nombreux problèmes se posent alors : comment choisir la taille de cet espace pour à la fois gagner en temps de calcul et ne pas trop perdre d'information, mais également quel noyau choisir pour représenter au mieux les données et proposer des techniques efficaces de comparaison.

DNA sequencing methods have shown significant technological advances in the past decades, giving access to a lot of data. High-throughput sequencing methods give access to a huge number of samples from which it is hard to perform efficient and fast classification tasks with usual methods. This is the case for example in metagenomics where millions of short sequences, called reads, have to be attributed to bacteria only based on their DNA sequences. Usual kernel methods such as SVM raise computational and storage issues making it impossible to scale them using the kernel trick. Other methods such as direct sequences' comparison are indeed efficient but impossible to scale given the high number of samples.
Kernel trick methods use the idea that every kernel, under certain hypotheses, can be viewed as a scalar product in a Hilbert space. Scalar products in this space can be rewritten using the kernel function. Kernel trick based methods use the interesting properties of this embedding space without explicitly working in it. New approaches proposed about ten years ago decrease drastically the computation time and storage required by approximating this embedding space by a low-dimension space in which usual linear methods would be performed with fast optimization techniques such as the stochastic gradient descent. Other issues are nevertheless raised such as the trade-off for the space-dimension : choosing a small-dimensional space would make the algorithm fast but not accurate enough whereas taking a high-dimensional one would be irrelevant in terms of computation. The choice of the kernel is also a bottleneck for the performance.

Bioinformatique Mines ParisTech & Institut Curie - *Paris*

Ugo TANIELIAN

9h30-10h30

User embeddings pour l'appariement multi-appareils

Data mining: lifetime value prediction and video game user classification

La publicité en ligne est une industrie cruciale pour les éditeurs et annonceurs. La possibilité de diffuser des annonces pertinentes aux utilisateurs est devenue un objectif nécessaire pour les entreprises de publicité. Cependant, avec le développement des téléphones intelligents et les tablettes, la majorité des utilisateurs possède plusieurs appareils. Cela se traduit par des entreprises ayant moins d'informations sur les utilisateurs au niveau de l'appareil. Pour effectuer un profilage de l'utilisateur sophistiqué, en particulier pour la publi-

Online advertising is a crucial industry for both publishers and advertisers. The ability to serve relevant ads to the users has become a necessary goal for advertising companies. However, with the development of smart phones and tablets, the majority of users owns several devices. This results in companies having to deal with weak user identities at a device level. To perform sophisticated user profiling especially for online advertising, it is important to link the same user across multiple devices.

cité en ligne, il est important de relier les mêmes utilisateurs sur plusieurs périphériques.
L'objectif d'ensemble de ce stage est de répondre à ce problème et d'identifier une même personne sur tous ces appareils. En apprenant des récents travaux réalisés à l'aide de l'algorithme word2vec (dû à T. Mikolov) et de ses applications à d'autres domaines, nous aimerions créer de nouvelles variables, les « users embeddings » (ou plongement d'utilisateurs) crées à partir de réseaux neuronaux. Ces variables nous permettront par la suite, d'améliorer notre algorithme de classification des paires d'utilisateurs.

The ultimate goal of this internship is to answer this problem and identify the same person across devices. Learning from the recent developments around the word2vec algorithm (due to T. Mikolov) and its applications to other fields, we would like to create new neural features based on user embeddings. These features would then enable us to improve our algorithm classifying pairs of users.

Criteo
Paris

10h30-11h30

Analyse et prédiction d'entrées dans des magasins à l'aide de machine learning

Analysis and prediction of store enterings with machine learning

Raymond CHAN

De nombreuses enseignes s'intéressent aux flux de clients concernant leurs boutiques : entre les personnes qui passent devant le magasin, celles qui s'arrêtent devant la vitrine, celles qui entrent effectivement dans la boutique, celles qui sont intéressées par certains rayons et produits, celles qui attendent en caisse... En particulier, connaître et anticiper le nombre de personnes effectives à l'intérieur du magasin tout au long de la journée permet à l'enseigne de mieux gérer son effectif.
À cette fin on place des caméras à l'intérieur des boutiques et à l'aide d'algorithmes d'analyse d'image, on peut compter le nombre de clients qui entrent et qui sortent du magasin. Les données de comptage sont alors stockées et on étudie l'influence des différents paramètres sur l'afflux de clients : jour de la semaine, heure de la journée, vacances, météo, température... Le but de ce stage est de déterminer les paramètres prépondérants sur la fréquentation des acheteurs, en particulier avec du machine learning. A terme, peut-être pouvoir prédire une semaine en avance le flot de clients durant la journée.

Numerous companies would like to know the flow of customers regarding their shops: the number of customers who walk past the store, the customers who stop in front of the shop display, those who go inside the store, those who are interested by specific sections or products, those who queue at checkout... Especially, knowing and forecasting the actual number of people within the shop would allow companies to better manage their labour.
To that end, we put cameras inside the store and with computer vision algorithms, we count the number of customers who go in and out of the store. Counting data is stored and we study the influence of different parameters on the flow of customers: day of the week, hour of the day, holidays, weather, temperature... The goal of this internship is to determine which features are decisive on the amount of purchasers, using machine learning. We might be able to predict the flow of customers during the day a week before.

Digeiz
Paris

11h30-12h30

Application d'apprentissage statistique pour la prédiction de pression de pore

Machine learning application for pore pressure prediction

Jiaxu LIU

La pression de pore est la pression à laquelle le fluide à l'intérieur des pores d'une roche se maintient à sa profondeur. C'est une des plus importantes mesures en amont dans le secteur pétrolier, car elle est indispensable pour bien diriger les opérations de forage et de cimentation. Néanmoins, la pression de pore est mesurée rarement directement. Par contre, normalement elle est dérivée par des ingénieurs d'interprétation à partir de plusieurs mesures, y compris les rayons gamma, la vélocité acoustique, la résistivité de la formation, etc. Bien qu'il y ait déjà des

Pore pressure is the pressure at which the fluid contained within the pore space of a rock is maintained at depth. It is one of the most important measures in upstream oil industry as it is indispensable for safely conducting drilling and cementing operations. However, pore pressure is seldom directly measured in practice. Instead, it is derived by interpretation engineers using different measures including Gamma ray, acoustic velocity, formation resistivity, etc. Although there are several proved methods developed in the past, it takes considerable amount of time

méthodes éprouvées développées avant, c'est très prenant pour les ingénieurs d'interprétation de déterminer la bonne méthode et les meilleurs paramètres pour chaque tâche.
L'objectif de ce stage est de construire un modèle d'apprentissage automatique en apprenant des interprétations humaines pour automatiser le processus de la prédiction de pression de pore. Tout d'abord, on agrège les données de plusieurs puits de différents champs pétrolifères et on choisit les données importantes en considérant leurs disponibilités et leurs corrélations avec la pression de pore. Puis un modèle est développé et examiné. Finalement, une application est construite afin de faciliter le travail des ingénieurs d'interprétation.

for interpretation engineers to choose the right method and parameters for each job.
The goal of this internship is to build a machine learning model to learn from human interpretations in order to save time by automatizing the pore pressure prediction process. We first aggregate the data from multiple wells in different oil fields, and choose important data channels based on their availability and their correlation with pore pressure. Then a model is developed and tested. Eventually, an application will be built for interpretation engineers to use.

Schlumberger
Beijing


Joséphine GAILLET

14h-15h

Management relation-client et stratégie après-vente d'un constructeur automobile : intégration de données d'usage Web dans le calcul des scores de clients

Customer Relationship Management and after sales strategy for a car manufacturer: integration of web data into client scoring models

Les stratégies marketing des constructeurs automobile reposent principalement sur la segmentation et le ciblage de leurs clients afin d'optimiser les investissements dédiés. L'enjeu est de mieux connaître les clients afin de diriger plus finement les actions marketing : c'est-à-dire adresser, au moment opportun, un contenu auquel le client sera réceptif.
Dans le cadre de sa stratégie marketing après-vente, un constructeur automobile cherche à évaluer la probabilité d'occurrence de certains événements : quels sont les clients les plus susceptibles de renouveler leur véhicule dans les 6 prochains mois ? Vont-ils monter en gamme, ou au contraire opter pour un véhicule moins onéreux ? Pour répondre à ce type de problématique, le constructeur automobile classe ses clients en leur attribuant une note : un « score ». Celui-ci est calculé à partir d'informations contenue dans les bases de données clients, utilisées comme variables.
Si les modèles de scores actuellement utilisés ont prouvé leur efficacité, un des objectifs de ce stage est d'en améliorer la performance en y intégrant des données supplémentaires, issues du web. Les données présentes sur internet fournissent une quantité d'information inimaginable. Les données de leads, par exemple, sont des données relatives aux contacts entre le constructeur et le client via internet, comme par exemple une demande d'e-brochure. Celles-ci étant déjà intégrées aux modèles de scores, nous serons amenés à évaluer leur impact sur la performance de scores. Nous tâcherons également d'aller plus loin en intégrant aux modèles de scores des données issues de nouvelles plateformes web, notamment des réseaux sociaux.

Car manufacturer marketing strategies are based on client segmentation and targeting in order to optimize investments dedicated to marketing. Client segmentation and targeting aim at better understanding who the clients are, in order to address them a relevant content, which they will be receptive to.
As part of its after sales marketing strategy, a car manufacturer is interested in evaluating the probability of occurrence of events such as: who are the clients most likely to renew their vehicle in the next 6 months? Will they upgrade or downgrade their vehicle? To assess the likelihood of such events, client scores are being calculated using data available into its client database.
Although actual scoring models have proven their efficiency to rank clients, the aim of this internship is to improve their performance by adding web data. Data available on the internet represent a massive amount of information. Lead data are already used in the scoring models. The leads give information about contact between the car manufacturer and its clients made through the internet, like an eBrochure request for example. The goal is to assess the impact lead data have on scoring model efficiency, and to go further by adding data from new Internet platforms, like social networks for instance.

Ekimetrics
Paris

15h-16h


Stéphane MENG

Détermination de modèles de prédiction d'avaries sur des organes donnés d'autobus

Estimation of fault prediction models on selected functions of a bus

Ce projet s'inscrit dans le cadre du projet Bus2025, qui vise à transformer l'ensemble du parc Autobus d'Île-de-France en 80% de bus électriques et 20% de bus au Biogaz, ainsi que de préparer l'ouverture à la concurrence des lignes de bus fin 2024. Actuellement, la maintenance du parc bus de la RATP comporte trois volets : maintenance préventive systématique, conditionnelle et la maintenance corrective. Le levier de compétitivité du réseau RATP repose sur une maintenance préventive systématique et une grande réactivité lors d'avaries, qui ont permis de générer un grand volume de données de maintenance.
L'objectif de ce stage est de commencer à exploiter ces données de maintenance en ciblant d'abord les organes les plus souvent défectueux, afin de déterminer des modèles de prédiction de panne. La première étape consiste à trier les données de maintenance et d'y intégrer les variables susceptibles d'expliquer les pannes. On essaie notamment de croiser des données de maintenance, d'exploitation et d'autres facteurs externes pouvant influer sur le comportement mécanique du bus. Ensuite, il s'agit de déterminer les modèles les plus adaptés pour prédire la possibilité des pannes futures. On pourra enfin essayer d'autres modèles à base de Chaînes de Markov par exemple pour tenter d'améliorer la maintenance corrective en déterminant la meilleure action à effectuer lors d'une panne future.

This project is part of the Bus2025 project, which aims at transforming the entire Ile-de-France bus fleet into 80% electric buses and 20% Biogaz buses as well as prepare the opening to the competition of bus lines at the end of 2024. Currently, the maintenance of the RATP bus fleet has three components: systematic preventive maintenance, conditional maintenance and corrective maintenance. The competitiveness lever of the RATP network is based on a systematic preventive maintenance and a high reactivity during faults, which allowed to generate a large volume of maintenance data.
The objective of this internship is to start using these maintenance data by first targeting the most often defective organs in order to determine fault prediction models. The first step will be to sort the maintenance data and integrate the variables that may explain the failures. In particular, an attempt will be made to compare maintenance, operating and other external factors that may affect the mechanical behavior of the bus. The next step is to determine the most appropriate models to predict the possibility of future failures. Finally, we can try other models based on Markov chains for example to try to improve the corrective maintenance by determining the best action to be carried out during a future failure.

RATP
Paris

Mercredi 28 juin *Wednesday 28th June* ■ V106A

8h30-9h30


Carole HARRY

Analyse et comparaison d'estimations du nombre de reproductions par différentes méthodes d'inférence

Analysing and comparing reproduction number estimates using different inferential methods

Le nombre de reproductions est un paramètre clé dans la l'étude de la transmissibilité d'un pathogène. Par définition, le taux de reproduction de base (R0) représente le nombre moyen d'individus qu'une personne infectieuse, dans une population entièrement susceptible, va contaminer au cours de son infection. Lorsque seule une partie de la population est susceptible d'être contaminée, le nombre de reproduction est appelé nombre de reproduction effectif (Re). Au cours d'une épidémie, l'efficacité de stratégies de contrôle peut être mesurée par l'évolution au cours du temps de la valeur du nombre de reproduction. Néanmoins, des études ont montré

The reproduction number is a key epidemiological parameter characterizing the transmissibility of a pathogen. By definition, the basic reproduction number (R0) is the average number of secondary infections generated by a typical infected individual in a fully susceptible population. When the population is only partially susceptible, the reproduction number is called effective reproduction number (Re). During the course of an epidemic, the effectiveness of a control strategy can be assessed by comparing estimates of the reproduction number computed at different times. Preliminary exploratory analysis suggests

des différences significatives, en fonction des méthodes utilisées, dans les estimations du nombre de reproduction et dans l'incertitude associée à ces estimations.
Ce stage vise à analyser les données issues de l'épidémie du virus Zika en Amérique latine. L'objectif est de recenser différentes méthodes d'estimation du nombre de reproduction, d'estimer le nombre de reproduction (et l'incertitude associée) pour Zika à partir de ces différentes méthodes, et de les comparer. Il est également prévu d'étudier les conditions (en particulier choix de la fenêtre de temps) sous lesquelles EpiEstim (un package implémenté dans R estimant le taux de reproduction), donne des résultats en adéquation avec les autres méthodes d'estimation.

that there could be significant differences in the reproduction number estimates (and uncertainty around them) depending on the method of estimation used.
In this project we propose to analyze the incidence data from the recent Zika epidemics in Latin America. The objective is to review different estimation methods, and apply them to Zika data to compare the corresponding reproduction number estimates (and their uncertainty). We also aim to explore the conditions, and specifically the time windows and time lags in EpiEstim (a statistical package R), under which the reproduction number estimates obtained with the various methods agree.

Imperal College
London, UK


Sébastien BILLEAU

9h30-10h30

Modélisation des prix de marché spot de l'électricité : une approche entre programmation linéaire et apprentissage statistique

Modelling electricity spot prices: a new approach mixing linear programming and statistical learning

Dans le contexte actuel de transition énergétique vers des moyens de production d'électricité renouvelables et du fait de la libéralisation du marché de l'électricité français, il est intéressant de modéliser le comportement des marchés spot de l'électricité et la manière dont ces derniers peuvent réagir à des changements du mix énergétique. L'actuelle sortie des tarifs d'achat pour les producteurs d'énergies renouvelables pour passer à un complément de rémunération pose la question de la place de nouveaux acteurs dans le marché de l'électricité, à savoir les agrégateurs. En effet, ce mécanisme impliquant une vente directe sur la bourse de l'électricité, les plus petits d'entre eux préféreront passer par ce nouvel acteur pour qu'il gère lui-même la vente d'électricité d'un grand nombre de petits producteurs. Modéliser les prix spot de l'électricité permet ainsi à ces petits producteurs de mieux gérer leur marge de négociations avec les agrégateurs.
L'objectif de ce stage est double. D'abord, il s'agit de développer une nouvelle méthode de modélisation des marchés spot de l'électricité qui mêle les deux types d'approches existantes à ce sujet, à savoir la programmation linéaire et l'apprentissage statistique. En effet, ces deux approches ont chacune des inconvénients qui rendent leur utilisation difficile dans une perspective prospective. Par ailleurs, il s'agit de rendre ce nouveau système utilisable dans le cas de la France, à la fois pour analyser ses performances sur les données des dernières années, mais aussi pour être fonctionnel vis-à-vis de la formulation de scénarios pour l'avenir.

The current energy transition towards renewable energies and the liberalisation of the French electricity market make modelling electricity spot market behaviour and response to any change in the energy mix particularly interesting. France is currently stopping its feed-in tariffs policy and replacing it by a remuneration compensation mechanism which raises the question of the role of new economic players, that is aggregators. Indeed, this compensation mechanism implies for energy producers to sell their production directly on the electricity exchange. Therefore, it is more convenient for small producers to have their production sold by an aggregator. Modelling electricity spot prices enables these small producers to better manage their negotiations with aggregators.
This internship has two main goals. First, a new method is developed to model electricity spot prices by mixing the two existing main methods, namely linear programming and statistical learning, as both methods have drawbacks that make them unsatisfactory in a prospective perspective. Furthermore, this new system has to be operational in the French case to analyse its performance with past data but also to be used as a predictive system for future scenarios.

Centre PERSEE, Mines ParisTech
Sophia-Antipolis

(PUBLIC RESTREINT/*RESTRICTED AUDIENCE*) 10h30-11h30

Maxime DEROUBAIX

Développement de modèles statistiques pour les risques climatiques et agricoles

Development of statistical models for climatic and agricultural risks

Les aléas climatiques peuvent avoir un impact significatif sur les résultats d'exploitation des entreprises. Outre le secteur de l'agriculture, dans lequel une carence d'ensoleillement ou un excès de précipitations ont une influence directe et évidente sur les récoltes, de nombreux domaines allant de l'énergie à l'alimentation sont également exposés aux anomalies météorologiques. Les exemples sont nombreux : chute de 15% des ventes de bière en 2012 en France suite à un été pluvieux, baisse de 20% de la quantité de gaz distribué en France lors de l'hiver 2014 particulièrement doux… Dans le but de permettre aux entreprises de transférer de tels risques, une solution a été développée : l'assurance paramétrique, qui repose sur la compensation financière du client suite à une irrégularité climatique mesurée par un paramètre. Les enjeux majeurs de la mise en place d'un tel produit d'assurance sont de trouver un paramètre fortement corrélé à l'activité de l'entreprise, de définir un seuil au-delà duquel l'assurance sera déclenchée, et d'établir le montant de la prime payée par le client en compensation du risque transféré. Pour dimensionner cette dernière, on utilisera dans un premier temps des méthodes (géo)statistiques pour modéliser le phénomène sur la base de données historiques, puis on lancera des simulations afin d'évaluer les pertes escomptées.

Climate hazards can have a substantial impact on company revenues. Beside the agriculture sector, for which a lack of sunlight or too much rainfall obviously affect yields, many industries such as energy or food are also exposed to weather anomalies. Examples are numerous: 15% drop in 2012 beer sales in France after a rainy summer, 20% collapse of gas distribution in France during the remarkably warm winter of 2014…
Parametric insurance has been developed to cover companies against such risks. A financial compensation is paid automatically to the client when an extreme weather event occurs. The main challenges of the product set-up are (i) to find a parameter highly correlated to the company's activity, (ii) to define a threshold beyond which an extreme weather event will be declared and the insurance triggered, and (iii) to determine the premium to be paid in compensation for the risk transfer. This premium will be dimensioned in two steps: we will first apply (geo)statistical methods on historical data to model the risk, and we will launch simulations to determine the expected loss.

AXA Global Parametrics
Paris

11h30-12h30

Maxime TONDEUR

Assimilation de données couplée sur un modèle couplé océan-atmosphère

Coupled data assimilation on a low-order coupled ocean-atmosphere model

L'assimilation de données est une technique pour combiner les prévisions d'un modèle et les observations. Elle a permis de réduire drastiquement les erreurs de prévisions en météorologie. Des solutions approchées ont été développées à partir de l'Ensemble Filtre de Kalman (EnKF) qui combinent une approche de Monte Carlo avec des hypothèses de linéarité et gaussiennes. Cependant, le couplage n'est pas présent avec les modèles actuels. La principale difficulté dans un modèle couplé est la différence d'échelle d'évolution de temps, ici l'océan est bien plus lent que l'atmosphère.
Le but de ce stage est d'étudier les propriétés du système d'assimilation de données quand il est couplé et qu'il s'applique sur un modèle couplé. Un modèle quasi-géostrophique qui décrit les températures et fonction de courant de l'océan et l'atmosphère développés dans une base de Fourier est utilisé. Des études de stabilité sont faites pour étudier les propriétés dynamiques de ce modèle. Pour

Data assimilation is a technique to combine the forecasts of a model and observations. It helps to reduce drastically the forecast error in meteorology. Approximate solutions are developed with the Ensemble Kalman Filter (EnKF) which combines a Monte Carlo approach with linearity and gaussianity hypothesis. Yet the coupling is not present in the current models. The main difficulty in a coupled model is the time-scale differences between ocean and atmosphere, here ocean quantities evolve more slowly than those of the atmosphere.
The goal of the internship is to study the properties of a data assimilation system when it is coupled and when it is applied to a coupled model. A quasi-geostrophic model which describes temperatures and stream functions of the ocean and the atmosphere expanded on a Fourier basis is used. Stability studies are done to study dynamical properties of this model. To achieve

cela les paramètres physiques de couplage sont changés pour se placer dans différents régimes. Puis il faut choisir un benchmark d'expériences d'assimilation de données en jouant sur la taille de l'ensemble et les observations par exemple. Le but est d'essayer de relier les propriétés dynamiques aux performances de l'assimilation de données et de comparer ces résultats aux résultats connus du cas non couplé.

it, coupling physical parameters are changed to test different regimes. Then a benchmark of data assimilation experiments is chosen by changing the size of the ensemble or the observations for example. The goal is to try to relate the dynamical properties to the performances of the data assimilation and to compare theses results to a known non-coupling case.

NERSC
Bergen Norvège
Institut Royal de Météorologie
Bruxelle , Belgique

Reporté/*Postponed* Septembre/*September*


Mélanie GITTARD

Comment expliquer les indicateurs de bien être dans les villes de l'OCDE à partir de données d'aménagement du territoire de Google Maps ?

How Google Map amenity datasets can explain well being in OECD cities ?

Estimer l'impact des politiques publiques d'aménagement des villes sur le bien être des citoyens est de plus en plus intéressant pour l 'étude de l'aménagement des territoires. Les progrès effectués pour collecter des données ont permis aux méthodes statistiques de devenir de nouveaux outils pertinents pour mesurer la qualité de vie des gens et aider à la décision politique. Les institutions internationales, telles que l'OCDE, utilisent ainsi l'analyse de ces données et le Big Data afin de créer de nouveaux indicateurs qualifiant le bien être, le « well-being », utilisés comme des références pour analyser et conseiller les politiques publiques.
Le but de ce stage, réalisé au sein des deux départements STD (département de statistiques) et GOV (département de la gouvernance publique et du développement territorial) de l'OCDE est de déterminer les regroupements d'aménités urbaines (comme les restaurants, cabinets médicaux, parcs…) qui vont avoir le plus d'impact sur la qualité de la vie dans les villes. Ce travail consiste à examiner les relations entre les méthodes d'aménagement des villes et la répartition des aménités utilisant des données scrappées depuis Google Maps et des indicateurs de bien-être, à construire durant le stage. Afin d'évaluer la qualité de ces données, la première étape du travail est de scrapper un autre site web d'aménités (comme les Pages Jaunes) et d'effectuer une validation croisée des deux bases. Est effectué ensuite le travail économétrique (ACP, BMA, modèle de régression…), afin d'identifier les types de services corrélées avec des données subjectives de bien-être tirées de l'enquête Gallup Analytics aux Etats-Unis. Un travail descriptif sur la densité des données spatiales dans les villes européennes est également mené.
Ce projet s'inscrit dans la devise de l'OCDE «Meilleures politiques pour une vie meilleure» : il vise la construction d'une mesure qui permettra d'aider les politiques urbaines à améliorer le bien-être de leurs citoyens.

There is a growing interest in land use planning to estimate the impact of urban public policies on citizen's well being. Thanks to the increase in the amount of data collected, statistical methods are relevant tools to measure the quality of people's life. Therefore data analysis and big data methods are used by International Institutions, such as OECD, in order to build new well-being indicators, considered as references to analyse and advise public policies.
The goal of this internship at both OECD STD (Statistics Directorate) and GOV (Directorate for Public Governance and Territorial Development) is to determine the mix of urban amenities (such as restaurants, doctors, parks…) that impact the most on city life quality. This work consists of looking at the relationships between spatial land use metrics and their distribution in cities using a scrapped dataset from Google Maps and indices of well being that have to be built. In order to assess the accuracy and completeness of Google Maps data, the first step is to scrap other websites such as the Yellow Pages and cross-validate the two datasets. Then, we undertake an econometric work (ACP, BMA, Regression model…) to identify the categories of amenities that impact and correlate the most with subjective well-being data drawn from Gallup Analytics survey in the United States. We also make a descriptive work of the density of amenities in Europe cities.
This project is in line with the OECD slogan « Better policies for better lives » : it aims to build a measure in order to help city policy planners to improve their citizen's wellbeing.

OECD
Paris

Etude d'un produit financier couvrant les entreprises pour les risques de change

Study of a financial product hedging companies against foreign exchange risks

Audrey MOUGNY-TANCHOU

La plupart des entreprises a besoin de se couvrir pour faire face aux différents risques qu'elles peuvent rencontrer. Dans le cadre de ma mission nous nous intéressons au risque de change, ce risque est lié à l'instabilité du taux de change d'une devise par rapport à une autre en fonction du temps. Les entreprises qui auraient besoin de se couvrir contre ce risque sont principalement les entreprises qui importent et/ou exportent. Au sein de la société générale, la BDDF (Banque de Détail De France) et la SG CIB (Corporate Investment Banking) ont créé une Joint-venture appelé MARK/FIC/NET qui aura pour but de couvrir les entreprises du risque de change. Au cours de cette mission, mon travail consiste à aider les inspecteurs dans la définition des axes à améliorer pour agrandir la clientèle, ceci en faisant une étude sur la clientèle (segmentation client, clustering…) et en inspectant la conformité des produits vendus.

Most of the companies need to hedge against the various risks that they can meet. In the framework of my mission, we are interested in the exchange rate risk, this risk is connected to the instability of the exchange rate of a currency compared with another one according to time. The companies which would need to hedge against this risk are mainly companies that export and import. Within the Société Générale, the BDDF (France's retail banking) and the SG CIB (Corporate Investment Banking) created a Joint-venture conscript MARK/FIC/NET which will aim to hedge companies of the exchange rate risk. During this mission, my work consists in helping inspectors in the definition of axes that we have to improve in order to enlarge the number of customers. To do this, I have to study my customer base (customer segmentation, clustering…) and verify the conformity of the sold products.

Société générale
Paris

L'influence des chocs bancaires et de la politique de financement sur l'innovation

Impact of banking shocks and financing policy on innovation

Nicolas JEANRENAUD

Pendant la Grande Dépression des années 30, la vague de faillites financières a eu un fort impact négatif sur l'accès au capital par le biais d'emprunts bancaires. Pourtant, malgré cette crise sans précédent, les données quantitatives montrent qu'il s'est agi de la décennie la plus innovante du vingtième siècle. Sous la direction de Professeur Tetyana Babina, ce projet cherche à comprendre comment l'innovation et la productivité peut être stimulées en temps de crise.
Les chocs bancaires chamboulent l'allocation des investissements indépendants et ont tendance à détruire l'accès au capital pour les petites entreprises, incitant les inventeurs et entrepreneurs à rejoindre des grands groupes. La concentration des talents qui s'en suit peut avoir des retombées positives sur l'innovation et la productivité, remettant en question la conception traditionnelle que l'innovation se fait au mieux dans les petites structures. L'objectif de cette recherche est d'analyser des données de brevets et de panels investisseurs-employeurs, des indicateurs locaux de détresse financière et des variables macroéconomiques, afin d'établir un lien entre santé du secteur financier, allocation des investissements et développement technologique, et aboutir à une meilleure compréhension des conditions favorables à l'innovation.

During the Great Depression of the 1930s, thousands of bank failures had a strong negative impact on access to capital through bank lending. Yet, despite this unprecedented level of financial distress, aggregate productivity statistics show that the decade was the most innovative of the twentieth century. Under the direction of Professor Tetyana Babina, this research project aims at understanding what drives innovation and productivity in times of crisis.
Bank financing shocks reshuffle the allocation of independent investment and can destroy access to capital for small firms, urging inventors and entrepreneurs to join larger companies. The consequent increase of talent concentration can lead to innovation and productivity spillovers, challenging the traditional view that innovation is best done in small firms. The goal of this research, through the analysis of patent data, investor-employer panel data, local financial distress indicators and macroeconomic variables, is to establish a link between financial sector health, investment reallocation and technological development, and reach a better understanding of the conditions that favour innovation..

Columbia Business School
New York, Etats-Unis