Ferdinand BHASVAR

Développement de méthodes de simulation géostatistique spatio-temporelle à l'aide des modèles génératifs du deep learning

Development of geostatistical spatio-temporal methods of simulations using generative models of deep learning

Résumé du projet de thèse: La géostatistique fournit des méthodes de prédiction d'une ou plusieurs variables en tout point d'un domaine, à partir d'observations parcellaires (mesures, sondages) voire indirectes (ex : données sismiques).

Quand l'objectif final n'est pas seulement la prédiction en chaque point mais également l'évaluation des incertitudes sur les résultats de calculs effectués sur la réalisation de la variable d'intérêt, on ne peut plus se contenter d'effectuer les calculs sur la base d'une prédiction, aussi bonne soit-elle, mais on doit connaître toute la distribution statistique des réalisations potentielles C'est-à-dire que l'on doit connaitre leur loi conditionnelle sachant les observations.

On a alors recours à des simulations stochastiques, conditionnées par les observations. La géostatistique fournit des méthodes de simulation efficaces basées sur des modèles parcimonieux. Avec l'accroissement de la quantité de mesures disponibles, les modèles géostatistiques classiques apparaissent parfois trop simples et on aimerait accroître leur réalisme.
Les modèles génératifs du deep learning, basés sur des réseaux de neurones, permettent de modéliser des phénomènes de manière beaucoup plus réaliste grâce à l'apprentissage de leurs paramètres sur de nombreuses images. Les methodes génératives les mieux établies sont les GAN (generative adversarial networks) et les VAE (variational autoencoders). Ces approches ont fourni des résultats spectaculaires dans différents domaines, de la biologie moléculaire à la génération d'images d'art.
L'objectif de la thèse est d'explorer le potentiel de ces méthodes pour remplacer, quand la quantité de données le justifie, les méthodes classiques de la géostatistique. De nombreux verrous nécessitent d'être levés pour y parvenir.

1) Apprendre des modèles dans le cas de structures stationnaires.
De cette façon, le modèle pourra être appris sur un petit bloc puis utilisé pour simuler des blocs beaucoup plus grands que ceux ayant servi à l'apprentissage.
2) Effectuer des simulations conditionnelles de ces modèles pour différents types d'observations (directes ou indirectes).

Les données directes sont des mesures du phénomène en certains points de l'espace tandis que les mesures indirectes sont des mesures produites par un modèle physique contrôlées par le phénomène.
Le candidat débutera la thèse en essayant de reproduire des blocs 3D générés par un logiciel de simulation de dépôts sédimentaires construit sur la base de la connaissance géologique des processus (Flumy).

Une fois cette étape réalisée, de nouvelles données environnementales, issues d'images satellites, seront utilisées. Dans cette optique, l'architecture des réseaux génératifs devra combiner spatial et temps. Les outils envisagés seront basés sur des réseaux de neurones récurrents (pour l'aspect temporel) et des réseaux convolutionnels (pour l'aspect spatial).

Dans un second temps, les efforts du candidat porteront sur le conditionnement par des observations (par inversion du réseaux de neurones génératifs). Ces observations seront tout d'abord directes, c'est-à-dire des points de mesure de la quantité à simuler observés en certains sites du domaine et certaines dates. Ensuite des observations plus complexes, issues par exemple d'un

modèle physique prenant en paramètre d'entrée la quantité inconnue à simuler (inversion bayésienne) pourront être utilisées.

Pour le conditionnement aux observations, quelque soit leur nature, l'approche envisagée sera basée sur le calcul bayésien variationnel et les résultats obtenus seront comparés à l'approche classique (mais plus coûteuse numériquement) par algorithmes MCMC.

Thesis abstract: Geo-statistic offers methods for predicting one or multiple variables in every points of a domain, from partial observations (measures, surveys) or even indirect observations (seismic data).

When the final objective is not only prediction in every points but also the evaluation of incertitude on the results of computation applied on the realization of the variable of interest, we cannot rely only on computations on one predictions, even a great one, but we need to know the whole statistical distribution of all possible realizations. This means we want to know their conditional law given the observations.

In this situation, we used stochastic simulations, conditioned with the observations. Geo-statistic offers efficient simulation methods based on parsimonious models. With the augmentation of the quantity of available measures, classical geo-statics models are often too simple and we would like to improve their realism.

Generative Deep-Learning models, based on Neural Networks, allow us to modelize phenomenon in a much more realistic fashion thanks to the training of their parameters on a huge number of pictures. Such models include GANs (generative adversarial networks) or VAEs (Variational Autoencoders). These approaches have proved to give formidable results, from molecular biology to generating art pictures.

The objective of the thesis is to explore the potential of these methods to replace, when justified by the number of data, classical methods of geo-statistic. Numerous obstacles have to be overcome :

1) Learn models in the stationary case. This way, the model can be trained on a small window of the data, but used to simulate bigger blocs.
2) Create conditioned simulations with these models for different types of observations (direct or indirect).
The candidate will start his thesis by trying to reproduce 3D blocs generated by a software of sedimentary deposit simulation based on geological processes (Flumy).

Once this step is done, new environmental data, from satellite imagery, will be used. In this context, the architecture of the generative networks will have to work both spatially and temporally. The deep learning tool considered for that step will be based on Recurrent Neural Networks (for time dimension) and convolution neural networks (for spatial dimensions).

In a second time, the efforts of the candidate will be focused on the conditioning given observations (through generative neural networks inversion). These observations will be direct first, in other words measuring points of the quantity of interest at some points of the domain and at certain dates. Then he will move on to more complex observations, for example coming from a physical model taking as input the unknown quantity to simulate (Bayesian inversion).

Regardless of the nature of the observations, the considered approach will be based on Variational Bayesian inference and the results will be compared to the classical approach (which is more costly computationally) through MCMC algorithms.