UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (ED 574)

Laboratoire Mathématique et informatique appliquées,
UMR 518 AgroParisTech-INRAE

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Pierre BARBILLON

Statistical Contribution to Uncertainty Quantification and the Analysis of Networks

Rapporteurs :

LILIANE BEL
CHARLES BOUVEYRON
DAVID HIGDON

Date de soutenance : 10 décembre 2020

Composition du jury :

CHRISTOPHE AMBROISE	(Examinateur)
LILIANE BEL	(Rapportrice)
CHARLES BOUVEYRON	(Rapporteur)
DAVID HIGDON	(Rapporteur)
JEAN-MICHEL MARIN	(Examinateur)
CATHERINE MATIAS	(Examinatrice)
OLIVIER ROUSTANT	(Examinateur)

Scientific production

Papers

Journal papers with methodological contribution

[JP1] P. Barbillon, L. Schwaller, S. Robin, A. Flachs, and G. D. Stone, *Epidemiologic network inference*, Statistics and Computing, pp. 1–15, 2019.

[JP2] M. Carmassi, P. Barbillon, M. Keller, Éric Parent, and M. Chiodetti, *Bayesian calibration of a numerical code for prediction*, Journal de la Société Française de Statistique, accepted, 2019.

[JP3] M. Courbariaux, P. Barbillon, L. Perreault, and É. Parent, *Post-processing multiensemble temperature and precipitation forecasts through an exchangeable normal-gamma model and its tobit extension*, Journal of Agricultural, Biological and Environmental Statistics, pp. 1–37, 2019.

[JP4] J. Ferrer-Savall, D. Franqueville, P. Barbillon, C. Benhamou, P. Durand, M.-L. Taupin, H. Monod, and J.-L. Drouet, *Sensitivity analysis of spatio-temporal models describing nitrogen transfers, transformations and losses at the landscape scale*, Environmental Modelling & Software, 111:pp. 356 – 367, 2019.

[JP5] T. Tabouy, P. Barbillon, and J. Chiquet, *Variational inference for stochastic block models from sampled data*, Journal of the American Statistical Association, (accepted), 2019.

[JP6] G. Damblin, P. Barbillon, M. Keller, A. Pasanisi, and E. Parent, *Adaptive numerical designs for the calibration of computer codes*, SIAM/ASA Journal on Uncertainty Quantification, 6(1):pp. 151–179, 2018.

[JP7] P. Barbillon, C. Barthélémy, and A. Samson, *Parameter estimation of complex mixed models based on meta-model approach*, Statistics and Computing, 27(4):pp. 1111–1128, 2017.

[JP8] P. Barbillon, S. Donnet, E. Lazega, and A. Bar-Hen, *Stochastic block models for multiplex networks: an application to a multilevel network of researchers*, Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(1):pp. 295–314, 2017.

[JP9] M. Courbariaux, P. Barbillon, and É. Parent, *Water flow probabilistic predictions based on a rainfall–runoff simulator: a two-regime model with variable selection*, Journal of Agricultural, Biological and Environmental Statistics, 22(2):pp. 194–219, 2017.

[JP10] G. Damblin, M. Keller, P. Barbillon, A. Pasanisi, and É. Parent, *Bayesian model selection for the validation of computer codes*, Quality and Reliability Engineering International, 32(6):pp. 2043–2054, 2016.

[JP11] E. Lazega, A. Bar-Hen, P. Barbillon, and S. Donnet, *Effects of competition on collective learning in advice networks*, Social Networks, 47:pp. 1–14, 2016.

[JP12] P. Barbillon, M. Thomas, I. Goldringer, F. Hospital, and S. Robin, *Network impact on persistence in a finite population dynamic diffusion model: application to an emergent seed exchange network*, Journal of theoretical biology, 365:pp. 365–376, 2015.

[JP13] M. Thomas, N. Verzelen, P. Barbillon, O. T. Coomes, S. Caillon, D. McKey, M. Elias, E. Garine, C. Raimond, E. Dounias *et al.*, *Chapter six-a network-based method to detect patterns of local crop biodiversity: Validation at the species and infra-species levels*, Advances in Ecological Research, 53:pp. 259–320, 2015.

[JP14] Y. Auffray, P. Barbillon, and J.-M. Marin, *Bounding rare event probabilities in computer experiments*, Computational Statistics & Data Analysis, 80:pp. 153–166, 2014.

[JP15] G. Damblin, M. Keller, A. Pasanisi, P. Barbillon, and E. Parent, *Approche décisionnelle bayésienne pour estimer une courbe de fragilité*, Journal de la Société Française de Statistique, 155(3):pp. 78–103, 2014.

[JP16] Y. Auffray, P. Barbillon, and J.-M. Marin, *Maximin design on non hypercube domains and kernel interpolation*, Statistics and Computing, 22(3):pp. 703–712, 2012.

[JP17] Y. Auffray, P. Barbillon, and J.-M. Marin, *Modèles réduits à partir d'expériences numériques*, Journal de la Société Française de Statistique, 152(1):pp. 89–102, 2011.

[JP18] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefebvre, and É. De Rocquigny, *Nonlinear methods for inverse statistical problems*, Computational Statistics & Data Analysis, 55(1):pp. 132–142, 2011.

Journal papers with consulting contribution

[JPC1] G. Carlin, C. Chaumontet, F. Blachier, P. Barbillon, N. Darcel, A. Blais, C. Delteil, F. M. Guillin, S. Blat, E. M. Van der Beek, A. Kodde, D. Tomé, and A.-M. Davila, *Maternal high-protein diet during pregnancy modifies rat offspring body weight and insulin signalling but not macronutrient preference in adulthood*, Nutrients, 11(1), 2019.

[JPC2] C. Lecarpentier, R. Barillot, E. Blanc, M. Abichou, I. Goldringer, P. Barbillon, J. Enjalbert, and B. Andrieu, *WALTer: a three-dimensional wheat model to study competition for light through the prediction of tillering dynamics*, Annals of botany, 123(6):pp. 961–975, 2019.

[JPC3] C. M. Bianchi, J.-F. Huneau, P. Barbillon, A. Lluch, M. Egnell, H. Fouillet, E. O. Verger, and F. Mariotti, *A clear trade-off exists between the theoretical efficiency and acceptability of dietary changes that improve nutrient adequacy* *during early pregnancy in french women: Combined data from simulated changes modeling and online assessment survey*, PloS one, 13(4):p. e0194764, 2018.

[JPC4] S. Fromentin, O. Davidenko, P. Barbillon, G. Fromentin, D. Tomé, and N. Darcel, *Variation in food preferences elicited by low-protein status in humans*, Appetite, 130:pp. 304–305, 2018.

[JPC5] M. Tharrey, F. Mariotti, A. Mashchak, P. Barbillon, M. Delattre, and G. E. Fraser, *Patterns of plant and animal protein intake are strongly associated with cardiovascular mortality: the adventist health study-2 cohort*, International journal of epidemiology, 2018.

[JPC6] O. Sauzet, C. Cammas, P. Barbillon, M.-P. Étienne, and D. Montagne, *Illuviation intensity and land use change: Quantification via micromorphological analysis*, Geoderma, 266:pp. 46 – 57, 2016.

[JPC7] J. Wencélius, M. Thomas, P. Barbillon, and E. Garine, *Inter-household variability and its effects on seed circulation networks. a case study from northern cameroon*, Ecology and Society, 2016.

[JPC8] C. Desclée de Maredsous, R. Oozeer, P. Barbillon, T. Mary-Huard, C. Delteil, F. Blachier, D. Tomé, E. van der Beek, and A. Davila, *High-protein exposure during gestation or lactation or after weaning has a period-specific signature on rat pup weight, adiposity, food intake, and glucose homeostasis up to 6 weeks of age.*, The Journal of Nutrition, (5), 2015.

[JPC9] A. Marsset-Baglieri, G. Fromentin, F. Nau, G. Airinei, J. Piedcoq, D. Rémond, P. Barbillon, R. Benamouzig, D. Tomé, and C. Gaudichon, *The satiating effects of eggs or cottage cheese are similar in healthy subjects despite differences in postprandial kinetics*, Appetite, 90:pp. 136 – 143, 2015.

Preprint

[P1] E. Baker, P. Barbillon, A. Fadikar, R. B. Gramacy, R. Herbei, D. Higdon, J. Huang, L. R. Johnson, A. Mondal, B. Pires, J. Sacks, and V. Sokolov, *Stochastic simulators: An overview with opportunities*, 2020.

[P2] S.-C. Chabert-Liddell, P. Barbillon, S. Donnet, and E. Lazega, *A stochastic block model for multilevel networks: Application to the sociology of organisations*, arXiv preprint arXiv:1910.10512, 2019.

[P3] K. Kamary, M. Keller, P. Barbillon, C. Goeury, and Éric Parent, *Computer code validation via mixture model estimation*, 2019.

[P4] M. Keller, G. Damblin, A. Pasanisi, M. Schuman, P. Barbillon, F. Ruggeri, and E. Parent, *Validation of a computer code for the energy consumption of a building, with application to optimal electric bill pricing*, 2019.

[P5] T. Tabouy, P. Barbillon, and J. Chiquet, *misssbm: An r package for handling missing values in the stochastic block model*, 2019.

[P6] A. Bar-Hen, P. Barbillon, and S. Donnet, *Block models for multipartite networks. applications in ecology and ethnobiology*, 2018.

[P7] M. Carmassi, P. Barbillon, M. Chiodetti, M. Keller, and E. Parent, *Calico: a r package for bayesian calibration*, 2018.

[P8] Y. Auffray and P. Barbillon, *Conditionally positive definite kernels: theoretical contribution, application to interpolation and approximation*, 2009, tech. Report.

In Preparation

[IP1] P. Barbillon, A. Forte, and R. Paulo, *Screening the discrepancy function*, (in preparation for submission).

[IP2] P. Barbillon and E. B. Pitman, *Embedding discrepancy within the computer model*, (to be continued).

Thesis

[T1] P. Barbillon, *Kernel interpolation methods for estimating expensive black box functions*, Ph.D. thesis, Université Paris Sud - Paris XI, 2010, URL `https://tel.archives-ouvertes.fr/tel-00559502`.

[T2] P. Barbillon, *Modèles réduits à partir d'expériences numériques*, Master's thesis, Université Paris Sud - Paris XI, 2007.

Editorial

[E1] A. Pasanisi, P. Barbillon, B. Iooss, and H. Monod, *Editorial of the special issue: Computer experiments, uncertainty and sensitivity analysis*, Journal de la Société Française de Statistique, 158(1):pp. 1–3, 2017.

Softwares

[R1] P. Barbillon and S. Donnet, *Gremlin*, `https://github.com/Demiperimetre/GREMLIN`, 2019.

[R2] T. Tabouy, P. Barbillon, and J. Chiquet, *misssbm*, `https://cran.r-project.org/web/packages/missSBM/index.html`, 2019.

Contents

Introduction 1

Introduction (version française)

Depuis mon doctorat, ma recherche s'est essentiellement concentrée sur des phénomènes complexes. J'ai considéré deux types de complexité qui correspondent aux deux chapitres suivants. Dans le chapitre 2, la complexité vient des modèles numériques, appelés aussi simulateurs, avec lesquels nous travaillons. Ces simulateurs permettent de faire des expériences dites numériques ou in silico. Elles remplacent les expériences physiques lorsque celles-ci ne sont pas réalisables ou sont trop coûteuses. Quand le phénomène modélisé est complexe, le simulateur doit prendre en compte de nombreux processus afin de décrire avec le plus de détails possibles les mécanismes. Il peut alors être coûteux en temps de calcul et dépendre d'un grand nombre de variables d'entrées. Dans de nombreuses situations, le simulateur est seulement disponible sous la forme d'une boîte noire, c'est-à-dire qu'il produit après un certain temps de calcul une sortie pour une configuration donnée en entrée, la séquence de calcul transformant les entrées en sortie est quant à elle inaccessible. La quantification des différentes sources potentielles d'incertitudes est une question primordiale lorsque l'on utilise un tel simulateur. Ce thème de la quantification des incertitudes pour les simulateurs a émergé récemment en statistique et est au cœur de certaines de mes contributions. Dans le chapitre 3, la complexité résulte de la structure particulière des données étudiées. Les données représentent les interactions entre des individus pouvant être des humains, des espèces animales ou végétales, des gènes, etc. et sont sous la forme de réseaux qui consistent en des ensembles de nœuds et d'arêtes reliant ces nœuds. Lorsque l'on modélise un réseau, la structure de dépendance est une source majeure de complexité. De plus, un réseau rend compte d'un type de relation spécifique entre certains individus mais d'autres types d'interaction peuvent exister et ces mêmes individus peuvent aussi interagir avec d'autres individus. Ce sont les réseaux multicouches qui se déclinent en réseaux multipartites, multiplexes et multiniveaux. L'observation partielle des réseaux due à un effort d'échantillonnage limité ajoute un niveau supplémentaire de complexité. Ces deux sources de complexité dues à un échantillonnage partiel et des réseaux multicouches sont au centre de mes contributions dans le domaine des réseaux.

Durant mon doctorat, j'ai commencé à travailler sur la quantification des incertitudes dans les simulateurs. J'ai alors continué à contribuer à cette thématique en tant que maître de conférence à AgroParisTech. J'ai notamment co-encadré avec É. Parent les thèses de G. Damblin (soutenue en 2015), M. Courbariaux (soutenue en 2017), M. Carmassi (soutenue en 2018) et je co-encadre actuellement avec J. Enjalbert la thèse de E. Blanc (soutenance prévue fin 2020). J'ai également collaboré avec deux collègues en post-doctorat J. Ferrer-Savall (2015-2016) and K. Kamary (2017). Le domaine de la quantification d'incertitudes est riche en collaborations potentielles puisque les simulateurs sont de plus en plus utilisés dans beaucoup de domaines scientifiques et sont très employés dans des applications industrielles. Les trois thèses soutenues étaient en collaboration avec trois département différents d'EDF (Électricité de France), respectivement : "modélisations des risques et des incertitudes", "production d'hydro-électricité" et "production d'électricité photovoltaïque". La thèse de M. Courbariaux était également en collaboration avec avec Hydro-Québec, un producteur d'énergie québécois. Le post-doctorat de K. Kamary était aussi en collaboration avec le département "modélisations des risques et des incertitudes" d'EDF. Ces collaborations avaient pour but de rendre compte des incertitudes dans la production d'énergie et d'évaluer la sécurité des centrales de production d'énergie. Le post-doctorat de J. Ferrer-Savall était en collaboration avec J.-L. Drouet de l'unité de recherche ECOSYS de l'INRA. Le sujet était la cascade de l'azote dans les territoires. La thèse d'E. Blanc est en collaboration avec l'unité de recherche du Moulon spécialisée en génétique végétale. Le but est d'étudier la croissante des plantes cultivées à l'aide d'un simulateur. Dans la communauté scientifique, la quantification d'incertitude dans les simulateurs est d'intérêt majeur comme le montre la formation de nombreux groupes de recherche réunissant mathématiciens appliqués, informaticiens des mondes académiques et industriels. Ces groupes ont été formés ces vingt dernières années : MASCOT NUM en France, MUCM au Royaume-Uni et un groupe activity group on UQ de la société SIAM (Society for Industrial and Applied Mathematics) aux États-Unis. Une revue intitulée Journal on UQ co-édité par l'ASA (American Statistical Association) et SIAM est même dédiée à cette thématique. Un numéro spécial [E1] du journal de la SFdS (Société Française de Statistique) portant sur cette thématique a également été publié en 2017. Des conférences internationales fréquentes ou des sessions spéciales dans les conférences internationales portent sur la quantification d'incertitude. Par exemple, SAMO (Sensitivity Analysis of Model Output), UQ16. L'institut de recherche SAMSI en Caroline du Nord aux États-Unis a organisé durant l'année universitaire 2018-2019 un programme de recherche MUMS (Model Uncertainty: Mathematical and Statistical) qui a réuni des chercheurs de tout horizon dans le domaine de la quantification d'incertitude. L'institut a accueilli des chercheurs visiteurs tout au long de l'année et a organisé des conférences au cours de l'année. J'ai eu la chance de participer à ce programme grâce au soutien financier d'une bourse Agreenskills+ pour un séjour à l'étranger.

J'ai commencé à travailler sur l'analyse statistique des réseaux en prenant part à un projet de recherche portant sur l'impact d'un réseau en tant qu'entrée d'un simulateur dynamique. Ce projet était en lien avec le post-doctorat de M. Thomas (2012) et la question de recherche était l'impact de la structuration sociale entre fermiers sur la biodiversité cultivée. Ces travaux continuent grâce au groupe de recherche MIRES (Méthodes Interdisciplinaires sur les Réseaux d'Échanges de Semences). Il a été financé en 2013 et en 2014 par le Réseau National des Systèmes Complexes et depuis 2015 par le département MIA de l'INRA. Il réunit des ethnobiologistes, des statisticiens, des géographes, des généticiens et des écologues. Je me suis ensuite intéressé aux différentes topologies de réseaux ce qui m'a apporté de nouvelles opportunités de travailler en particulier sur les topologies dérivant de modèles à blocs, ces blocs représentant des groupes de nœuds. Avec mes collaborateurs, nous avons ensuite étendu les modèles à blocs à des réseaux multicouches et avons proposé des méthodes d'inférence adaptées. Avec S. Donnet, nous co-encadrons S.-C. Chabert-Liddell sur ces questions (soutenance prévue fin 2021). Le traitement de l'inférence de modèles à blocs en présence de données manquantes dans les réseaux d'interaction a été traité dans la thèse de T.

Tabouy (soutenue en 2019) que j'ai co-encadrée avec J. Chiquet. De plus, l'inférence d'un réseau en tant que paramètre d'un simulateur dynamique a été l'une de mes contributions. L'analyse des réseaux est intéressante pour de nombreuses applications notamment en sociologie et en écologie. En sociologie, nous avons collaboré avec E. Lazega qui s'intéresse à la sociologie des organisations et qui a pour but de déterminer comment les relations entre les individus et les relations entre les entreprises s'entremêlent. En écologie, nous collaborons avec des écologues principalement grâce à un financement ANR Econet qui a commencé en 2019. Le but de ce projet est de développer des méthodes statistiques afin de répondre spécifiquement à l'analyse des différents types de réseaux écologiques (réseaux trophiques, réseaux hôtes-parasites, plantes-pollinisateurs, plantes-champignons) et de comprendre les mécanismes qui guident les interactions entre espèces.

De nombreux modèles dans mes contributions, autant dans le chapitre 2 que dans le chapitre 3 comportent des variables latentes. J'utilise de tels modèles dans le chapitre 2 soit pour décrire grâce à la variable latente les différents régimes d'erreur du simulateur ou pour modéliser des effets individus dans un modèle à effets mixtes. Les variables latentes dans le chapitre 3 correspondent aux regroupements en blocs qui structurent les interactions dans les réseaux. J'ai utilisé des techniques d'inférence qui sont soit bayésienne, soit qui reposent sur un algorithme type Espérance-Maximisation (EM) ce qui est assez proche dans le domaine des modèles à variables latentes. Cela peut nécessiter de simuler les variables latentes (dans un cadre bayésien ou lorsque l'on utilise des versions stochastiques de l'algorithme EM) et les paramètres à estimer (dans le cadre bayésien). Pour ce faire, j'utilise des méthodes MCMC notamment les algorithmes de Gibbs et de Metropolis-Hastings. En alternative aux versions stochastiques de l'algorithme EM, j'utilise une version variationnelle de l'algorithme EM dans le chapitre 3 où le calcul exact de l'étape E est prohibitif et est remplacé par une approximation variationnelle qui rend les calculs réalisables. En plus de ces techniques, le cœur des outils utilisés dans le chapitre 2 repose sur la théorie des processus gaussiens qui permettent d'obtenir des émulateurs du simulateur.

Les deux chapitres suivants posent un contexte introductif nécessaire pour présenter mes contributions. Bien qu'écrits en anglais, ils présentent un résumé long en français. Dans le chapitre 2, les contributions portent sur des méthodes d'analyse de sensibilité de simulateurs de grande dimension, sur le calage et la validation de simulateurs et sur la modélisation du régime d'erreur du simulateur. Je présente ensuite mes perspectives poursuivant les travaux menés notamment sur la modélisation des erreurs de simulateurs ou s'ouvrant vers de nouvelles thématiques comme les simulateurs stochastiques. Dans le chapitre 3, l'organisation est quelque peu différente. Les notations et le vocabulaire communs sont tout d'abord présentés. Ensuite, les contributions sont séparées en trois parties thématiques : influence du réseau dans un modèle complexe, inférence d'un réseau de contact et modélisation par des modèles à blocs de réseaux d'interaction. Elles sont exposées avec un contexte introductif propre suivi des éléments principaux de la ou des contributions correspondantes. Finalement, les perspectives de ce chapitre sont regroupées dans une partie finale. Plusieurs d'entre elles sont notamment liées à des questions propres aux réseaux écologiques d'interaction.

Introduction (English version)

Since obtaining my Ph. D., my research contributions have been mainly focused on complex phenomena. I have considered two kinds of complexity which correspond to the following two chapters. In Chapter 2, complexity comes from the computer models a.k.a. simulators that we are dealing with. These simulators are used to run numerical or computer experiments. They replace field experiments when unfeasible or too costly. When the modeled phenomenon is complex, the simulator has to account for many processes to be accurate. It may then become time consuming and depend on many input variables. In most situations, the simulator is only available as a black box i.e. it provides an output for a given input configuration after a certain amount of time but the detailed sequence of computations transforming the inputs into the output is unreachable. It is often necessary to assess the different sources of uncertainties when working with such simulators. This is the aim of the area of Statistics called "Uncertainty Quantification" in which I have several contributions. In Chapter 3, complexity arises as a result of the particular structure of the data at hand. The data represents interactions between individuals (humans, species, gene, etc.) and are given as networks which consist of sets of nodes and sets of edges linking these nodes. When modeling a network, the structure of dependencies is an important source of complexity. Moreover, a network accounts for some specific relations between specific individuals but other kinds of interactions may exist and the same individuals may be implied in other interactions with other individuals. This then results in multilayer networks which are of different types: multiplex, multipartite, multilevel to name but a few. The partial observation due to a limited sampling effort adds another layer of complexity. My main contributions in the network area deal with this complexity induced by limited sampling and multilayer networks.

My Ph.D. was concerned with some aspects of UQ. I then continue to contribute to this area as an assistant professor at AgroParisTech. I co-supervised with É. Parent the Ph.D. theses of G. Damblin (defended in 2015), M. Courbariaux (defended in 2017), M. Carmassi (defended in 2018) and I co-supervise with J. Enjalbert the Ph. D. thesis of E. Blanc (defense planned in late 2020). I also collaborated with two post-doctoral fellows J. Ferrer-Savall (2015-2016) and K. Kamary (2017). This area of research is rich in potential collaboration since the use of simulator is more and more popular in many scientific domains and is widespread in many industrial applications. The three defended Ph. D. theses were carried out in collaboration with three different departments of EDF (Électricité de France), respectively: "uncertainties and risks modeling", "hydro-electricity production" and "photovoltaic electricity production". The Ph. D. thesis of M. Courbariaux was also in collaboration with Hydro-Québec, a Québec energy supplier. The post-doctoral fellowship of K. Kamary is also in collaboration with the department "uncertainties and risks modeling" of EDF. These collaboration with EDF were dedicated to quantify the uncertainties in energy production and to assess the safety of powerplants. The post-doctoral fellowship of J. Ferrer-Savall was achieved in collaboration with J.-L. Drouet from the INRA ECOSYS research unit and was concerned with the cascade of nitrogen in the landscape. The Ph. Thesis of E. Blanc is in collaboration with Le Moulon, a research unit in plant genetics. The goal is to study the plant growth through a simulator. In the scientific community, uncertainty quantification in simulators is of major interest as demonstrated by the formation of numerous research groups aggregating applied mathematicians, computer scientist, statisticians from academic and industrial teams. These groups were formed in the last decades: Mascot

NUM in France, MUCM in the UK, and SIAM (Society for Industrial and Applied Mathematics) activity group on UQ in the USA. A joint ASA (American Statistical Association) /SIAM Journal on UQ is devoted to papers in this field. A special issue [E1] of the Journal of the SFdS (Société Française de Statistique) was also published in 2017 on this subject. Regular international conferences or special sessions in international conferences on UQ are numerous and popular. To name a few: SAMO (Sensitivity Analysis of Model Output, UQ16. SAMSI a research institute in North Carolina, USA organized during the academic year 2018-2019 a program entitled MUMS (Model Uncertainty: Mathematical and Statistical) which gathered international researchers in the field of UQ. The institute hosted research visitors throughout the year and organized several research workshops along the year. I was part of this program thanks to the financial support of an Agreenskills+ outgoing fellowship.

I began to work on the statistical analysis of networks by being part of a research project where the question was to study the impact of a network as an input of a dynamic simulator. This project took place during the post-doctoral fellowship of M. Thomas (2012) and was concerned with the impact of the social structure of farmers on the cultivated biodiversity. These works keep going on through the research group MIRES (Méthodes Interdisciplinaires sur les Réseaux d'Échanges de Semences). It was funded in 2013 and 2014 by the Réseau National des Systèmes Complexes and since 2015 by the INRA department MIA. It brings together ethnobiologists, statisticians, geographers, geneticists and ecologists. I then focused on the different topologies of networks. This brought new opportunities to work on the inference of particular topologies based on a blockmodeling of the network (clustering of nodes). With my collaborators, we then extended the blockmodels to multilayer networks and proposed a dedicated inference method. With S. Donnet, we supervise S.-C. Chabert-Liddell on these questions (defense due in late 2021). The question of dealing with missing data in the inference of blockmodels was central in the Ph. Thesis of T. Tabouy (defended in 2019) which I supervised with J. Chiquet. Furthermore, the question of inferring the network which serves as an input of a dynamic simulator was investigated. The analysis of network is interesting for many applications including sociology and ecology. In sociology, we collaborated with E. Lazega who focuses on the sociology of organizations and aims to determine how relations between individuals and between organizations are intertwined. In ecology, we are collaborating with ecologists mainly through an ANR grant named Econet which started in 2019. The goal is to develop statistical methods specifically designed for analyzing different types of ecological networks (food webs, host-parasite, plant-pollinator, plant-fungus networks). This will allow us to understand the mechanisms that determine species interactions.

Many models in my contributions, as well in Chapter 2 as in Chapter 3, include latent variables. I resort to latent variable modeling in Chapter 2 either to describe through the latent variables different error regimes for the simulator or to represent individual effect in mixed effect model. The latent variables in Chapter 3 correspond to the block clustering which shapes the interactions in networks. The inference techniques I used are mainly either Bayesian or rely on Expectation-Maximization (EM) algorithm. Technically, the algorithms are quite close in the area of latent variable models. They may require to simulate latent variables (in a Bayesian framework or when using a stochastic version of the EM algorithm) and parameters to be estimated (in a Bayesian framework). I used MCMC algorithms including Gibbs and Metropolis-Hastings algorithms. As an alternative to stochastic versions of the EM algorithm, I used variational version of the EM in Chapter 3 where the exact computation of the E step is replaced with a variational approximation making the computations tractable. In addition to these techniques, the core of the techniques in Chapter 2 are Gaussian processes which produce surrogates (a.k.a. metamodels) of simulators.

The following two chapters provide an introductory context necessary for presenting my contributions. In the chapter 2, the contributions relate to methods of sensitivity analysis of simulators in high dimension, to the calibration and validation of simulators and to the modeling of the simulator error regime. Then, I present my perspectives, either continuing the work carried out in particular on the modeling of simulator errors or opening up to new themes such as stochastic simulators. In the chapter 3, the organization is somewhat different. The common notations and terminology are first presented. Then, the contributions separated into three parts: influence of the network in a complex model, inference of a contact network and blockmodeling interaction networks, are exposed with their own introductory context followed by the main elements of the contribution(s) in question. Finally, the perspectives of this chapter are gathered in a final part. Several perspectives are notably linked to questions specific to ecological networks of interaction.

Uncertainty Quantification 2

Résumé du chapitre en français

Les simulateurs sont des implémentations de modèles mathématiques de phénomènes réels complexes. Ils ont une importance cruciale dans de nombreux champs scientifiques. Un appel à un simulateur (une simulation) pour une configuration d'entrées choisie est appelé une expérience numérique ou in silico. Ces expériences remplacent les expériences de terrain ou en laboratoire lorsque celles-ci sont trop coûteuses en ressources ou ne sont pas réalisables. La quantification d'incertitude s'attache à modéliser et prendre en compte les différentes sources d'incertitude lorsque l'on travaille avec un simulateur. L'incertitude entache certains paramètres d'entrées du simulateur qui doivent être choisis par l'utilisateur avant de lancer une simulation. En confrontant les sorties du simulateur à quelques expériences de terrain, il est possible de réduire cette incertitude. Cette tâche s'appelle le calage du simulateur. Cette confrontation permet également de mesurer l'écart entre le simulateur et le phénomène réel qui est appelé la discrépance. Quantifier cette discrépance permet de valider ou non le simulateur, c'est-à-dire décider si le simulateur est assez précis au regard de son utilisation prévue. Les simulateurs sont généralement des fonctions boîte-noire dans le sens où ils ne sont pas disponibles sous forme analytique. Ils sont seulement disponibles comme un code exécutable renvoyant une sortie pour une entrée. Ce code est le plus souvent coûteux en temps de calcul, un appel pouvant durer plusieurs heures voire des jours. Il est alors souvent nécessaire d'employer des techniques de réduction de modèle afin de construire un émulateur qui sera une version rapide et approchée du simulateur. Cet émulateur est construit à partir d'un nombre limité d'appels bien choisis au simulateur. Travailler avec un émulateur ajoute alors une couche d'incertitude supplémentaire.

Mes contributions dans ce domaine peuvent être regroupées en quatre parties principales : i) l'analyse de sensibilité pour les simulateurs en grande dimension, ii) les problèmes inverses dont le calage, iii) l'erreur de simulateur et iv) le post-traitement de prévisions hydrologiques et météorologiques.

L'analyse de sensibilité a pour but d'identifier quelles entrées du simulateur ont le plus fort impact sur ses sorties. Cela permet de mieux comprendre le phénomène modélisé ainsi que de simplifier l'utilisation du simulateur en limitant la dimension des entrées. Bien que les méthodes d'analyse de sensibilité utilisées soient simples, la difficulté a été de gérer plusieurs types d'entrées qui sont spatialisées et temporelles. Les sorties sont également spatialisées et/ou dynamiques. Nous avons proposé dans [JP4] des méthodes de visualisation efficaces afin de résumer une grande quantité de sorties. J'ai principalement travaillé avec deux simulateurs, l'un modélisant la diffusion de l'azote dans un territoire agricole et l'autre modélisant la croissance de plants de blé.

Les problèmes inverses ont pour but d'estimer certaines entrées, appelées paramètres, du simulateur à partir d'expériences de terrain correspondant à des sorties du simulateur. L'estimation de ces paramètres s'appellent le calage. Dans [JP6], nous avons proposé un plan d'expérience numérique séquentiel adapté à l'objectif de calage. Dans le cadre d'un modèle mixte impliquant le simulateur, le paramètre d'entrée est supposé être tiré aléatoirement et indépendamment pour chaque individu d'une population. Le but est alors d'estimer les paramètres de la loi de la population. Dans [JP7], nous avons montré comment estimer ces paramètres tout en tenant compte de l'incertitude due à l'utilisation d'un émulateur à la place d'un simulateur trop coûteux.

Dans [JP10, P3], nous proposons de traiter la question de la validation d'un simulateur comme un problème de choix de modèle entre un modèle intégrant seulement un bruit d'observation contre un modèle comportant également un terme de discrépance.

Les articles [JP3] et [JP9] se concentrent sur le post-traitement de prévisions hydrologiques et météorologiques. Des simulateurs sont utilisés pour proposer des prévisions d'ensemble qui ont souvent besoin d'être post-traitées afin d'être utilisées en tant que prévision probabiliste. Nous avons proposé de les post-traiter en les intégrant dans un modèle statistique. Nous avons de plus proposé un modèle statistique avec deux régimes d'erreur pour post-traiter les sorties d'un simulateur pluie-débit.

Dans mes perspectives, la poursuite des travaux concernant la discrépance du simulateur s'articule autour de trois axes. La détection des variables auxquelles la discrépance est la plus sensible grâce à une sélection de modèle par calcul du facteur de Bayes devrait permettre de mieux comprendre les erreurs du simulateur. Des tests entre plusieurs modèles de discrépance offrent également une possibilité de mieux comprendre l'erreur du simulateur et d'améliorer sa prise en compte dans des prédictions. Enfin, lorsque le simulateur dérive d'équations différentielles, il est possible d'incorporer directement un terme de discrépance dans ces équations. Cela pourrait permettre une meilleure prise en compte des incertitudes, directement dans les parties incertaines du simulateur plutôt que comme un terme externe correcteur.

Bien que la plupart des simulateurs soient déterministes, les simulateurs stochastiques sont de plus en plus populaires. Les techniques usuelles de quantification des incertitudes doivent être alors étendues aux simulateurs stochastiques ce qui représente des difficultés supplémentaires. Avec des collègues, nous avons écrit un article de revue identifiant ce qui est fait et ce qui reste à faire [P1] dans ce domaine.

Les deux simulateurs pour lesquels nous avons effectué une analyse de sensibilité sont sources de défis méthodologiques majeurs qui donnent lieu à des perspectives pour des travaux futurs. Le simulateur modélisant la diffusion de l'azote dans le territoire est construit comme un couplage dynamique de quatre simulateurs. Émuler un tel simulateur demande de développer une méthodologie dédiée qui rendra possible l'exploration plus poussée de celui-ci. Pour le simulateur modélisant la croissance de plants de blé, la question est de réussir à construire une analyse de sensibilité pour un mélange de deux variétés de blés présentant des traits phénotypiques différents. Cela demande de proposer un plan de simulation adapté pour mesurer comment ce mélange peut conduire à de meilleurs rendements.

La question du choix des plans d'expériences numériques reste une question importante qui doit toujours être reliée au(x) objectif(s) visé(s). Notamment dans le contexte de la validation du simulateur, trouver les expériences numériques qui permettraient de détecter où le simulateur présente des écarts plus importants à la réalité permettrait de déterminer précisément un domaine de validité du simulateur dans un premier temps puis à l’améliorer dans un deuxième temps. Cette question se pose aussi pour le choix des expériences de terrain quand un choix est possible.

Since most of the contributions share a common background, the first section is devoted to the exposition of Uncertainty Quantification. This section deals with the general techniques and goals in this field. It also introduces some specific terminology. Then, the contributions and the perspectives are presented.

2.1 BACKGROUND ON UQ

2.1.1 General Overview

Simulators (a.k.a. computer models, mechanistic models, etc.) are mathematical models of complex real-world processes. They are a crucial ingredient in most fields of science, engineering, medicine and business. Running the simulator at a chosen set of inputs on a computer is sometimes named a numerical experiment or an experiment in silico. These experiments replace wet lab experiments or physical experiments when they are too costly or impracticable. For instance, simulators in biology/medicine describe load decrease in HIV patients or tumor growth [175, 78, 106, 141]; in pharmacokinetics, compartment models are used to simulate the absorption and elimination of a drug dose given to a patient [JP7]; in agronomy, the coupling of hydrological, atmospheric, land use and agricultural simulators models the cascade of nitrogen at the landscape level [60, JP4]; in nuclear safety, simulators allow engineers to assess the reliability of a nuclear reactor in specific working conditions and its acess to cooling source of water [P3, 27]; in energy production, simulators help to predict the production of a power plant [JP10, JP2]; in meteorology, simulators provide forecasts of future meteorological conditions [JP9, JP3]; in natural hazard quantification, hazard maps are derived from simulators to inform public policies [155, 160]; in socioeconomics land use and transportation integrated models help to evaluate planning policies and development scenarios [71]. In the last decades, the growth in computing power increased the usage of simulators and they were devoted to explore larger and larger problems. From a statistical point of view, the main issue when working with simulators is uncertainty quantification (UQ). Uncertainties are present at different levels of the simulators. Uncertainties may affect the inputs of the simulator. Some inputs of the simulator are called parameters and are not precisely known. Generally, only a coarse information is available such as a reference value and lower and upper bounds. This uncertainty on input parameters can be modeled as a probability distribution function in a Bayesian framework. In spite of the complexity of the simulator, it can suffer from a discrepancy with real experiments. This discrepancy should be carefully evaluated to assess or not the validity of replacing physical experiments with simulator runs. Moreover, the simulator is time consuming since a call may take hours or even days. In this case, model reduction techniques are necessary to explore the set of inputs and to study its behavior. This limited time budget leads to an additional source of uncertainty.

The uncertainty quantification task consists in taking into account all these sources of variability and to determine to what extent the simulator is reliable to describe the real-world process. The goal is also to reduce these uncertainties by using all available data which are from heterogeneous sources. Statisticians have proposed Sensitivity Analysis (SA) [146] methods or screening techniques [111] to determine which inputs have the most impact on the outputs of the simulator. By doing so, the input dimension can be reduced to the space of the influent input variables. When the simulator is too costly in computation time, model reduction techniques (a.k.a. emulation[149]) which incorporate also information on the approximation quality are proposed. These techniques launch a limited number of runs of the simulator to build a fast approximation of the simulator over the whole set of inputs. Choosing the designs of numerical experiments, i.e. input configurations where the simulator is run, is also a key question for statisticians. Since the computational budget is limited, this design has to be chosen carefully with respect to the intended goal. A general requirement for the design of numerical experiment is to be space filling [62, 47, JP16]. Since the emulators such as GPE are interpolators, their performances depend on the coverage of the input space by the design of experiments. The designs can be enriched sequentially when a specific goal is fixed such as optimizing the simulator [89] or computing a probability of failure corresponding to the exceedance of a threshold by the output of the simulator [9]. When field experiments (experiments generated from the real-world process) are available, the unknown parameters of the simulators can be calibrated. The simulator is to be embedded into a statistical model and the calibration is dealt with as an inference problem [93]. The simulator running time, the large dimension of inputs and outputs and the limited number of field data make the calibration a challenging problem. Eventually, the simulator has to be validated which means that it has to be demonstrated that it sufficiently well represents the real-world process. Validation task is not simply answering the naive question: "Can the simulator represent adequately the reality ?" It should be context dependent. After assessing the uncertainty affecting the simulator, the simulator is declared valid or not regarding its intended uses [8]. These steps are summarized in Figure 2.1. After the problem specification where the experts identify the inputs of the simulator: input variables $\mathbf{x}$ and parameters θ with a corresponding probability distribution representing the uncertainties (Step 1), a sensitivity analysis or a screening technique is run (Step 2) in order to focus only on the most influential inputs (input variables and/or parameters) in the following steps. If running the simulator f is time consuming, an emulator is to be constructed from a design of numerical experiments (Step 2bis). From a cheap simulator or from an emulator, optimization, visualization and computation of exceedance probabilities [JP14] can be performed. Calibration and validation tasks (Steps 3 and 4) are based on the simulator or its emulator and on additional field data (not to be confounded with numerical experiments which consists of runs of the simulator). Calibration consists in reducing the uncertainties on input parameters θ of the simulator and validation consists in determining whether or not the simulator is sufficiently close to the real-world process f^R.

Figure 2.1 omits some possible feedback loops. These feedback loops intervene when the design is built sequentially as mentioned above for optimization or estimation of a failure probability. Sequential numerical design of numerical experiments may also enhance the calibration [JP6].

On the top of this general presentation, some particular features of simulators should be identified and properly taken into account since they need specific developments. The simulator is said to be deterministic or stochastic whether two runs of the simulator at the same inputs provide the same outputs or not. Stochasticity may be the result of two typical cases. First, stochastic approximation schemes such as Monte Carlo computations or more sophisticated schemes [84] are used in the computations of the simulator and lead to different output for different random seeds. Second, the simulators coming from fields such as biology, ecology, epidemiology, socioeconomics embrace a model for stochasticity. Agent-Based Models

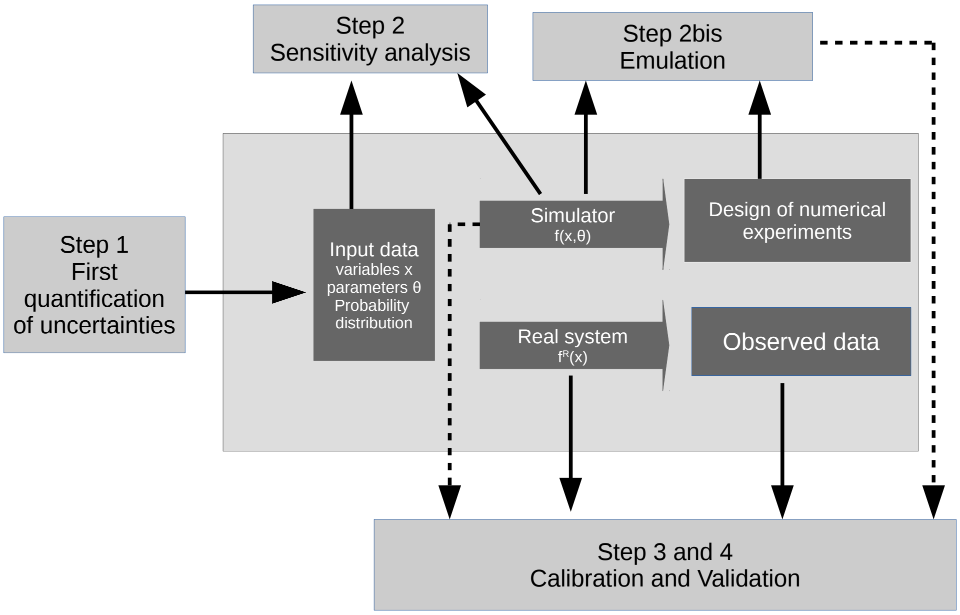


Figure 2.1 – *Global approach in Uncertainty quantification*

(ABM) [70] (see `https://www.comses.net/` and `https://ccl.northwestern.edu/netlogo/references.shtml` for online platforms dedicated to ABMs) also known in ecology as Individual-Based Models (IBMs) are a particular case of this kind of model where individuals (agent) are modeled in interaction with others. Since the possible interactions are complex, they are often modeled as stochastic. For instance, in epidemiology, the spread of a disease can be modeled as an ABM with an SIS/SIR (Susceptible Infected Susceptible / Recovered) model. Individuals have random interactions with one another and may or may not be contaminated by an infected ones. In this case, the whole distribution of the output for a given input configuration is of interest. Indeed, the quantile of the spread of the epidemics could help to assess the level of risk. Conversely, when the stochasticity results from numerical schemes, only the mean is the quantity of interest. The variability around the mean is seen as noise that can be reduced by replicating the calls to the simulator at the same input configuration.

The input and/or output of a simulator may be functional. For instance in an environmental simulator, the practitioner needs to specify among the inputs the maps of used soil and the climatic sequence with the temperature and the amount of rain [JP4]. Some outputs are spatially and temporally distributed as the emission of some chemical constituents for each pixel of the spatial map and for each day of the simulated period. To deal with these high dimensional inputs and outputs, the main practical solutions consist in reducing the dimension thanks to a projection method such as Principal Components Analysis (PCA) or Singular Value Decomposition (SVD) and/or functional basis such as polynomial basis or the Karhunen-Loève expansion [68] which is appropriated to represent a stochastic process. The usual techniques in UQ rely on these lower dimensional approximation to deal with these cases of functional outputs (see [85] for calibration and [7]). For emulating simulator with time-series output, another proposition is to use time-series statistical models where some stochastic process are assumed to be Gaussian processes (GP) over the simulator input space [113]. The works with functional input are scarcer [87] but this is currently an active field of research.

The computer model may consist of several simulators coupled in the sense that the inputs of a simulator are the outputs of another simulator. One can consider the global computer model as a unique simulator in which case the analysis and the emulation are performed by the usual techniques. However, some recent works [100, 119] show that one can take advantage of uncoupling the simulators to emulate them separately and then couple the emulator. When the same simulator is self-coupled iteratively meaning that the outputs of the simulator obtained at a given step become the inputs of the simulator for the next step and so on, a solution could be to emulate the simulator which models the transition between two steps instead of emulating the global simulator which outputs all the series of outputs at all steps [40].

2.1.2 Statistical Models and Notations

In this section, we introduce the main notations and the terminology used in the next sections. Although these terms were informally introduced above, we give more precise definitions of the key concepts. We also introduce some classical statistical models that incorporate the simulator and which are common to some contributions and perspectives.

Simulator. The simulator is a function denoted by f. It is often a black-box function in the sense that for a given input a computer code produces an output but the function has no analytical expression or its expression is unknown to the practitioner. In some papers, the term “computer code” actually means the simulator. This code may consist in a numerical solver of ODEs or PDEs. An evaluation of the simulator is time consuming, which makes any repetitive task involving the simulator challenging. This computation cost is generally translated into a maximum number of runs of the simulator which is called the simulation budget.

Its input variables may be functional as well as its outputs but for the sake of simplicity, we assume that $f : \mathbb{R}^q \to \mathbb{R}$. Furthermore, we assume that the inputs are only continuous variables and not discrete or qualitative which should require specific considerations [176]. We denote by $\mathbf{z} \in \mathbb{R}^q$ or by $(\mathbf{x}, \theta) \in \mathbb{R}^p \times \mathbb{R}^d$ the inputs depending on whether it is necessary or not to make a distinction between the parameters θ of the simulators and the input variables $\mathbf{x}$. The latter notation is used when physical / field experiments corresponding to this simulator are available. In this case, the $\mathbf{x}$'s have a physical counterpart and are observed while the θ's are unknown. The $\mathbf{x}$'s may be called controlled or environmental variables depending whether they can be set up by a user in the physical experiments while environmental variables can be measured but not controlled. For instance, controlled variable may be the speed and the altitude of an aircraft and environmental variables may be the outside temperature and humidity rate. Still, we will not make the distinction in the following since they can be treated the same way in our contributions. The parameters θ's may have a physical meaning such as physical constant (gravity acceleration constant,...) or just tuning parameters of the code. Again we will not make a distinction although the prior distributions may be more informative in the former case. When we use the notation $\mathbf{z}$, it means that the inputs may be of any kind or a mixture of variables and parameters but the distinction is not relevant for the given task. For instance, when performing a sensitivity analysis or emulating all the inputs can be processed in the same way.

Designs of experiments. Since the number of evaluations of the simulator is limited, the evaluations have to be carefully chosen with respect to the intended goals. The set of input configurations where the simulator is run is generally called the design of experiments. To emphasize that it corresponds to simulations, we will call it the design of numerical experiments (DoNE). It is available as a $N \times q$ matrix D where N is the number of evaluations of the simulator and has to be smaller than the computational budget. The rows of D are then the different chosen input configurations. Contrary to some real experiments, the practitioner is totally free to choose the input configurations. We denote by $f(D) = \{f(\mathbf{z}_1), \ldots, f(\mathbf{z}_N)\}$ the set of simulations for the input configurations that are in the DoNE. A desirable feature for a DoNE is to well cover the domain of interest which is generally a bounded set $B \subset \mathbb{R}^q$ and even given as a Cartesian product of interval for each dimension, i.e. $B = \prod_{j=1}^{q}[l_j, u_j]$. The good coverage property is also called the space-filling property. Many criteria [62, 47] are given as possible definitions of this property. One popular criterion is maximin criterion [88, JP16]: a design D is maximin if it maximizes the minimal distance between its pairs of points:

$$\min_{\mathbf{z},\mathbf{z}' \in D} ||\mathbf{z} - \mathbf{z}'||. \tag{2.1}$$

On the top of this coverage property, some projection properties can be met if the design is sought in the class of Latin hypercube design (LHD) [125]. An LHD has the property that the orthogonal projection of the points on any input dimension j should be such that there is one and only one point in every subinterval $[l_j + (u_j - l_j) \cdot k/N, l_j + (u_j - l_j) \cdot (k+1)/N)$ $(k = 0, \ldots, N-1)$ resulting from an equal subdivision of the domain interval with respect to this dimension.

To relate the DoNE with the intended goals, it is built in two steps: a first initial design is chosen to be space-filling and then new points are added to the design conditionally to the available information. More precisely, it consists in iterating from a given DoNE with n points: D_n, the steps:

1. Optimize a criterion Crit with respect to the current available information to find the new point: $\mathbf{z}_{n+1} = \arg\max_{\mathbf{z}} \text{Crit}(\mathbf{z}|f(D_n))$.
2. Evaluate $f(\mathbf{z}_{n+1})$, and complete the sets $D_{n+1} = D_n \cup \{\mathbf{z}_{n+1}\}$ and $f(D_{n+1}) = f(D_n) \cup \{f(\mathbf{z}_{n+1})\}$.

This adaptive approach corresponds to Bayesian optimization [123] and is particular well suited when a GP emulator (GPE) is used for f. The criterion Crit has to be adapted to the intended goals [89, 9, JP6]. To leverage the parallelization of runs, a batch of new points can be chosen and added to the design [37].

We can also have access to a design of field / physical experiments (DoFE) when the goal is to calibrate or validate the simulator. In this case, we have n_e recorded couples: $(\mathbf{x}_i^e, y_i^e)_{1 \le i \le n_e}$. The vectors $\mathbf{x}_i^e \in \mathbb{R}^p$ correspond to the input variables of the simulator. We recall that these p input variables are a subset of the inputs of the simulator which may also have as inputs some additional parameters not observed in the field experiments (in our notation $p \le q$). We use the $n_e \times p$ matrix notation D^e to aggregate the vectors $\mathbf{x}_i^e$ and the vector notation $\mathbf{y}^e$ for the set of observations which corresponds to the quantity of interest modeled by the simulator. The DoFE may be chosen or just observed depending on the case at hand. When the input variables are controlled, the question of choosing the experiments to run is relevant [133]. Otherwise when they are environmental, the only option is to measure them. We use the index or superscript e to refer to the field experimental data instead of f to avoid confusion with the simulator.

Emulator. To alleviate the computational burden due to a simulator, an emulator is built on the basis on its evaluations at the input configurations aggregated in the DoNE. An emulator is a fast approximation of the simulator which can also be named a surrogate model, a meta-model or a response surface. Two desirable features of an emulator are the interpolation on the points of the DoNE and the quantification of the error due to the replacement of the simulator. The interpolation makes sense when the simulator is deterministic since another run of the simulator at the same input configurations will produce the same output. The quantification of the additional uncertainty due to the emulator may be available as an additional variability which is easy to integrate in a statistical model. GP emulator (GPE) are popular emulators since they have these two features and are simple to manipulate. They derive from Kriging which comes from spatial statistics [98, 120, 44] and were first used to emulate a simulator in the seminal paper of Sacks et al. [145]. For the last decades, they were widely used and some connections have been made with kernel interpolation [151, 170] in approximation theory [T1].

To build an emulator for f, it is assumed to be a realization of a GP F over the space $\mathbb{R}^q$. Another interpretation is to consider the GP F as a prior distribution on the function f. The distribution of the GP is given as $F \sim \mathcal{GP}(m(\cdot), \sigma_F^2 C(\cdot,\cdot))$ where m is the mean function, σ_F^2 the variance and C the correlation kernel. The mean function is usually given as $m(\mathbf{z}) = \sum_{j=1}^m \beta_j h_j(\mathbf{z}) = H(\mathbf{z})^T \beta$ with h_j known functions such as linear or polynomial functions of the inputs and β_j's unknown parameters to be estimated. The m dimensional vector $H(\mathbf{z})$ contains the m functions evaluated in $\mathbf{z}$ and β aggregates the β's. The variance is also a parameter to be estimated. The correlation kernel is a symmetric positive definite function which depends on additional parameters such as the range with respect to the different inputs. Usually, the correlation kernel is assumed to be a product of univariate radial basis function (RBF) k, i.e. $C(\mathbf{z}, \mathbf{z}') = \prod_{j=1}^q k(|z_j - z_j'|)$. This implies that the Gaussian process is second order stationary.

The function k may be a power exponential function: $k(d) = \exp(-d^\alpha/\psi)$ where $\psi \in \mathbb{R}_+^*$ is a range parameter determining the distances for which the correlation still matters. The parameter α is a regularity parameter lying in $(0, 2]$. If $\alpha = 1$, k is called an exponential RBF and if $\alpha = 2$, k is called squared exponential RBF or Gaussian RBF. This class of RBF leads to GP continuous in mean square but (infinitely) differentiable in mean square only for $\alpha = 2$. Another popular RBF is the Matérn class of RBF, the general expression of which is rather technical since it depends on the Gamma function and a version of the Bessel function. It also depends on a regularity parameter denoted by ν and a range parameter also denoted by ψ. For $\nu = p + 1/2$ with $p \in \mathbb{N}$, we have simple analytic expressions. For example, when $\nu = 1/2$, we get the exponential RBF, for $\nu = 3/2$ $k(d) = (1 + \sqrt{3}d/\psi)\exp(-\sqrt{3}d/\psi)$, for $\nu = 5/2$ $k(d) = (1 + \sqrt{5}d/\psi + 5d^2/(3\psi^2))\exp(-\sqrt{5}d/\psi)$ and for $\nu \to \infty$ we retrieve the squared exponential RBF. The GP with a Matérn RBF is k-times mean square differentiable if and only if $\nu > k$. For more details on correlation kernels, see [140]. The range parameter ψ can be the same for any input dimension or indexed by the dimension ψ_j leading to q parameters. These two cases are respectively called isotropic or anisotropic. These range parameters ψ's as well as the β's in the mean function and the variance σ_F^2 have to be estimated whereas the regularity parameter and the class of RBF are fixed by the practitioner from some prior knowledge on the simulator f. Different approaches exist to deal with these hyperparameters leading to more or less computational burden. In a full Bayesian approach, prior distributions must be set. They can be also estimated by maximum likelihood estimates (MLE) from the data $f(D)$ and be then plugged into the GP. Instead of the MLE, the posterior modes of these hyperparameters can be plugged in. The use of specific prior distributions [77] helps to make the emulation more robust in the sense that it avoids the degeneracy of the covariance matrix computed at the DoNE locations. Cross validation methods can also produce estimates for this hyperparameters [5].

When the hyperparameters are fixed, the GP process F conditioned to the data $f(D)$ is still a Gaussian process with analytic expression for the mean and the variance:

$$\begin{array}{rcl} F|f(D) & \sim & \mathcal{GP}(m_{D,\beta}(\cdot), C_{D,\beta}(\cdot,\cdot)), \\ m_{D,\beta}(\mathbf{z}) & = & H(\mathbf{z})^T\beta + \Sigma_{\mathbf{z}D}^T\Sigma_D^{-1}(f(D) - H_D\beta) \\ C_{D,\beta}(\mathbf{z},\mathbf{z}') & = & \sigma_F^2(C(\mathbf{z},\mathbf{z}') - \Sigma_{\mathbf{z}'D}^T\Sigma_D^{-1}\Sigma_{\mathbf{z}D}^T) \end{array} \tag{2.2}$$

where H_D is a $N \times m$ matrix such that $(H_D)_{ij} = h_j(\mathbf{z}_i)$, $\Sigma_{\mathbf{z}D}$ is a N dimensional vector s.t. $(\Sigma_{\mathbf{z}D})_i = C(\mathbf{z}_i, \mathbf{z})$ and Σ_D is a $N \times N$ vector s.t. $(\Sigma_D)_{ij} = C(\mathbf{z}_i, \mathbf{z}_j)$ and $f(D)$ is considered as a N dimensional vector. If we choose a flat prior on $\beta \propto 1$, the conditioned GP is still Gaussian with analytic expression for the mean and the variance:

$$\begin{array}{rcl} F|f(D) & \sim & \mathcal{GP}(m_D(\cdot), C_D(\cdot,\cdot)), \\ m_D(\mathbf{z}) & = & H(\mathbf{z})^T\hat{\beta} + \Sigma_{\mathbf{z}D}^T\Sigma_D^{-1}(f(D) - H_D\hat{\beta}) \\ C_D(\mathbf{z},\mathbf{z}') & = & \sigma_F^2(C(\mathbf{z},\mathbf{z}') + u(\mathbf{z}')^T(H_D^T\Sigma_D^{-1}H_D)^{-1}u(\mathbf{z}) - \Sigma_{\mathbf{z}'D}^T\Sigma_D^{-1}\Sigma_{\mathbf{z}D}^T) \end{array} \tag{2.3}$$

with $u(\mathbf{z}) = H_D^T\Sigma_D^{-1}\Sigma_{\mathbf{z}D} - H(\mathbf{z})$ and $\hat{\beta} = (H_D^T\Sigma_D^{-1}H_D)^{-1}H_D^T\Sigma_D^{-1}f(D)$ the generalized least square estimator. For particular prior distributions on the couple (β, σ_F^2) it is still possible to obtain a known posterior process (Student process) but if a prior distribution is also considered for the correlation kernel parameters, the full distribution can be only sampled [149]. Since integrating the uncertainties on the β's is simple, the popular Gaussian process emulator comes from Equation 2.3. The conditional mean acts as an approximation of f over the domain of interest and has the property to interpolate the simulator (i.e. $m_D(\mathbf{z}) = f(\mathbf{z})$ if $\mathbf{z} \in D$). The conditional variance is a measure of uncertainty on the quality of the approximation. It is zero for the evaluations that have already been run, i.e. $C_D(\mathbf{z},\mathbf{z}) = 0$ if $\mathbf{z} \in D$. Figure 2.2 displays a simulator with one dimensional input space and its GPE from a design of 5 points. The simulator f is only known to the practitioner on 5 locations, a GP distribution with a constant mean and a Matérn 5/2 kernel for the correlation. The conditional mean and the pointwise 95% credibility interval for any location are provided. The conditional mean interpolates the 5 evaluations of f and the credibility interval tends to be narrower near the design locations up to be exactly zero at these locations.

Remark. If the simulator is stochastic, the covariance of the GP can be chosen as:

$$\mathrm{Cov}(F(\mathbf{z}), F(\mathbf{z}')) = \sigma_F^2 C(\cdot,\cdot) + \tau^2\delta_{\mathbf{z}=\mathbf{z}'}$$

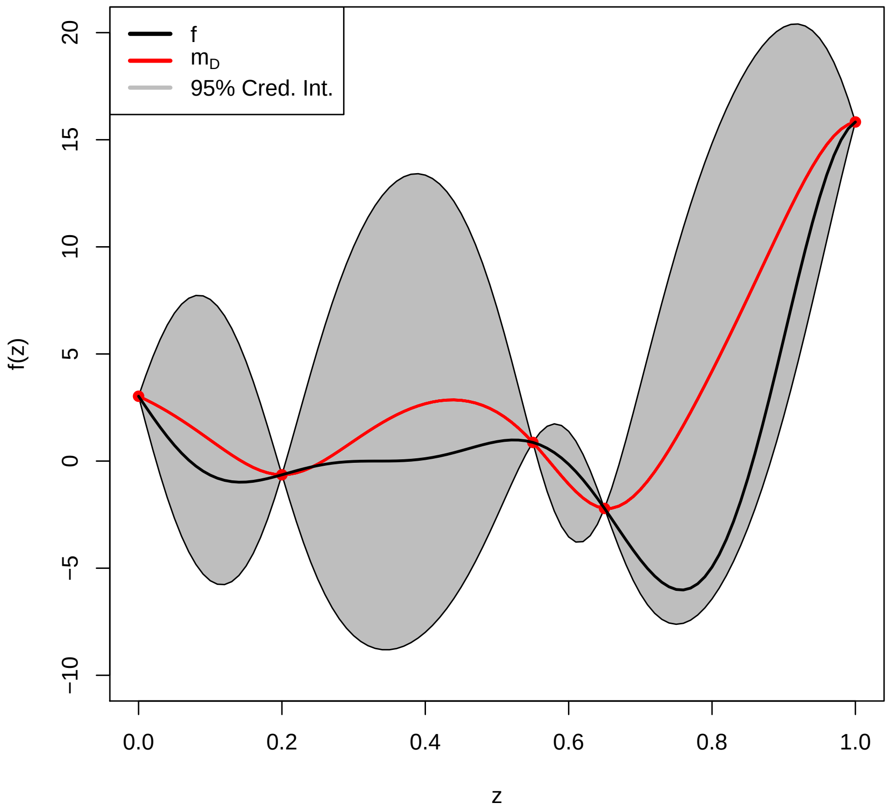


Figure 2.2 – *Example of a GPE from 5 evaluations (red dots) of the simulator f. The pointwise credibility interval is obtained from the conditioned variance.*

where $\delta_{\mathbf{z}=\mathbf{z}'} = 1$ if $\mathbf{z} = \mathbf{z}'$, $\delta_{\mathbf{z}=\mathbf{z}'} = 0$ otherwise. As a heritage from the spatial statistics literature, this additional term is called the nugget. It prevents the conditioned mean m_D from being an interpolator and the uncertainty at any point of the design is then τ^2. It can also act as a regularization parameter and leads to an emulator with better statistical properties, such as predictive accuracy and coverage, in a variety of common situations [75]. More sophisticated models are also possible for stochastic simulator. The nugget can also dependent on the inputs and be modeled as another GP [73, 14].

We can interpret the GPE as a kernel interpolator since a Reproducing Kernel Hilbert Space (RKHS) $\mathscr{H}_C$ can be associated to C the correlation kernel as soon as it is positive definite [3, 140]. We define $g(\mathbf{z}) = f(\mathbf{z}) - H(\mathbf{z})^T \hat{\beta}$ as the function to interpolate where $\hat{\beta}$ can be fixed or estimated by the general least square estimator. This function g is assumed to lie in the RKHS $\mathscr{H}_C$ and we consider the interpolation problem on D:

$$\begin{cases} \min_{v \in \mathscr{H}_K} ||v||_{\mathscr{H}_C} \\ \text{such that } g(\mathbf{z}_i) = v(\mathbf{z}_i),\ i = 1, \dots N. \end{cases}$$

The solution to this problem $\tilde{v}$ is equal to the second term in the expression of $m_{D,\hat{\beta}}$ in Equation (2.2) i.e. $\tilde{v}(\mathbf{z}) + H(\mathbf{z})^T \hat{\beta} = m_{D,\hat{\beta}}(\mathbf{z})$ for any $\mathbf{z}$ [152, T1] with $\hat{\beta}$ being plugged in. Moreover, this interpolator comes with a control of the pointwise error which is:

$$\forall z, |f(\mathbf{z}) - m_{D,\hat{\beta}}(\mathbf{z})| = |g(\mathbf{z}) - \tilde{v}(\mathbf{z})| \leq ||g||_{\mathscr{H}_C} P_D(\mathbf{z}),$$

where $P_D(\mathbf{z}) = C_{D,\hat{\beta}}(\mathbf{z},\mathbf{z})/\sigma_F^2$ from Equation (2.2). This bound being deterministic, the kernel interpolation vision is often used in theoretical proofs which requires a consistency of the emulator [JP16, JP7].

Screening and sensitivity analysis. The goal of screening and sensitivity analysis (SA) is to assess the impact of the inputs of the simulator on its output. We focus on global methods which assess the effect of an input over its domain of variation. Other methods are said to be local and provides the effect of making a small change for an input at a precise location.

Screening is a coarser method in the sense that it mainly separates inputs with no effect or a negligible effect from inputs which do impact the output. These two groups of variables are called respectively inert and active variables. A major method for screening is the Morris method [124] which consists in evaluating the effects of elementary displacements in a normalized input space on the output. Thus, a particular DoNE is related to this method which is called one-at-time since the elementary displacements are made with respect to each dimension successively the other ones being fixed. Another method is linked with the GPE [111] since the screening is performed on the parameters related to the inputs of the emulator.

SA methods provide a quantification of the effects of the inputs and help to order them from the most impactful to the least one. They can also provide information on the interactions between inputs. The basic methods for SA derive from linear models: linear regression and analysis of variance (ANOVA). The latter needs a discretization of the input space but is able to deal with a mixture of quantitative and qualitative inputs. The extension of the ANOVA model is the functional ANOVA coming from the Hoeffding decomposition which allows to define the so-called Sobol' indices [157].

Recent works [45] propose to use some dependence measure such as Hilbert-Schmidt independence criterion (HSIC) to consider the impact of an input on the whole distribution of the output and not only on the variance as the Sobol' indices do.

Calibration and validation. For performing calibration and validation, a DoFE and the corresponding results of field experiments must be available. These experiments and the simulator are related through a statistical model as proposed in the seminal paper of Kennedy and O'Hagan [93]. First, we assume that the field experiments are noisy observations of the real phenomenon f^R: for $i = 1, \dots, n_e$,

$$y_i^e = f^R(\mathbf{x}_i^e) + \epsilon_i \tag{2.4}$$

where a simple distribution can be assumed on the noise such as $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ since they are considered as measurement errors. Moreover some strong prior information can be available on σ_ϵ^2 since the precision of the measuring devices may be known. Second, a link between the real phenomenon and the simulator is assumed:

$$f^R(\mathbf{x}) = f(\mathbf{x}, \theta) + \delta(\mathbf{x}). \tag{2.5}$$

It then results in

$$y_i^e = f(\mathbf{x}, \theta) + \delta(\mathbf{x}) + \epsilon_i. \tag{2.6}$$

Recall that we use the notation $(\mathbf{x}, \theta)$ for the set of inputs since we want to distinguish between input variables which are observed on field experiments and input parameters which do not have a physical counterpart. Ideally, if the simulator was a perfect representation of the reality, it would exist a parameter θ^* such that $f^R(\mathbf{x}) = f(\mathbf{x}, \theta^*)$. The parameter θ^* is then interpreted as the true and the best fitting parameter. Unfortunately, it is rarely the case. Hence, a so-called discrepancy function or bias function δ is added to compensate for what the simulator misses. Obviously this parametrization leads to an identifiability issue since for two different couples (θ_1, δ_1) and (θ_2, δ_2) we may have $\delta_1(\mathbf{x}) = f^R(\mathbf{x}) - f(\mathbf{x}, \theta_1)$ and $\delta_2(\mathbf{x}) = f^R(\mathbf{x}) - f(\mathbf{x}, \theta_2)$. In a Bayesian perspective [93], the solution is to set a prior distribution on the discrepancy: $\delta \sim \mathcal{GP}(m_\delta(\cdot), \sigma_\delta^2 C_\delta(\cdot, \cdot))$. A common choice for the mean is $m_\delta = 0$ since the parameter θ is sought as the best fitting parameter which makes f as close as possible to f^R and then the discrepancy helps to compensate a systematic error. This systematic error depends on the input variables since it is expected that if the simulator has a flaw in its modeling for a given $\mathbf{x}$, this should be also the case in the neighborhood of $\mathbf{x}$. However, this GP prior for the discrepancy does not prevent some confounding effects with θ. Although the predictions issued by this model are generally good in spite of the confounding effects, the confounding is still puzzling for many statisticians and practitioners (see the discussions in [93]). A case for better specifications of the prior distribution of δ relying on expert knowledge was made in [25]. Other solutions to limit the confounding is to impose constraints on the GP modeling the discrepancy. The prior distribution on the discrepancy can be assumed to be orthogonal to the gradient of the simulator [137]. Another solution is to model the discrepancy by a scaled GP which consists in setting a non increasing prior distribution on the L_2 norm of the GP [76]. This will encourage the discrepancy to be as small as possible. If the prior distribution is flat it corresponds to the classical GP modeling for the discrepancy.

In a frequentist context, the model is similar but the estimation is done in two steps. First the parameter θ^* is estimated as the solution to $\arg\min_\theta \sum_{i=1}^{n_e} (y_i^e - f(\mathbf{x}_i^e, \theta))^2$.

Second, the data $(\mathbf{x}_i^e, y_i^e - f(\mathbf{x}_i^e, \theta^*))_{1\leq i \leq n_e}$ are used to learn δ with non-parametric estimates [166, 165, 173]. These methods can be referred to as L_2 calibration.

In the same spirit in the Bayesian literature, this approach which consists in splitting the inference in separated tasks is called modularization [112]. The general model is cut into different modules where the inference is conducted with the other being fixed. As in the L_2 calibration, the discrepancy module and the calibration parameters may be inferred in two steps. Moreover, in a joint inference without modularization, the field data should impact inference of the emulator although these data are of different nature than the evaluations of the simulator. The modularization for the emulator consists in using only the evaluations $f(D)$ to build it. This not only limits the computational burden but also prevents a flawed model for field data to contaminate the emulator. A comparison of the different practical solutions of [93, 86, 8] to conduct calibration is done in G. Damblin's Ph.D. Thesis [46].

There exist other techniques for calibration which are more focused on determining a subset of the input parameter space coherent with the observed field data rather than deriving a posterior distribution of the parameter. History matching [43, 172] accounts for the different sources of uncertainty and proceeds by exclusion of the regions of the input parameter space which are implausible with the field data. The bound to bound approach [65] deploys semidefinite programming algorithms where the initial bounds on calibration parameters are combined with initial bounds of experimental data to produce new uncertainty bounds for the calibration parameters that are consistent with the data.

The validation method proposed in [8] consists in providing tolerance bounds around the posterior predictive mean which should contain with a high probability the true real process. These bounds are computed by integrating the different sources of uncertainties on the simulator due to its emulation and its discrepancy with the real world process, its parameters, field data... The prediction for a new input variable location $\mathbf{x}_{new}$ can be either a pure-simulator prediction given as

$$\hat{f}(\mathbf{x}_{new}, \hat{\theta}) = m_D(\mathbf{x}_{new}, \hat{\theta})$$

where $\hat{\theta}$ may refer to the posterior mean or the posterior mode and m_D is used instead of f if we consider an expensive simulator, or a bias-corrected (discrepancy-corrected) prediction given as the mean:

$$\hat{f}^R(\mathbf{x}_{new}) = \frac{1}{M}\sum_{j=1}^{M}\Big(F^{(j)}(\mathbf{x}_{new}, \theta^{(j)}) + \delta^{(j)}(\mathbf{x}_{new})\Big)$$

where $F^{(j)}$ are posterior realizations of the GP F given $f(D)$ (evaluations of f at the DoNE) and $(\theta^{(j)}, \delta^{(j)})$ are sampled from the joint posterior predictive distribution deriving from Equation (2.5) given the field data $\mathbf{y}^e$. For a fixed level γ, the tolerance bounds $\tau = \tau(\mathbf{x})$ are then computed such that $\gamma \cdot 100\%$ of the samples satisfy:

$$\left|\hat{f}(\mathbf{x}_{new}, \hat{\theta}) - m_D(\mathbf{x}_{new}, \hat{\theta})\right| < \tau$$

for the pure-simulator prediction. Similarly for the bias-corrected prediction, τ are computed such that $\gamma \cdot 100\%$ of the samples satisfy:

$$\left|\hat{f}^R(\mathbf{x}_{new}) - \big(F^{(j)}(\mathbf{x}_{new}, \theta^{(j)}) + \delta^{(j)}(\mathbf{x}_{new})\big)\right| < \tau$$

A practitioner will then decide whether to use or not the posterior prediction issued by the simulator by comparing the width of the tolerance bounds with the accuracy required for the intended use. In our contributions [JP10, P3], we rather consider the validation task as hypothesis testing or as model selection problem where the two models in competition assume $\delta = 0$ or $\delta \neq 0$. This approach shares similarities with [132]. The authors propose to keep some field data on which a validity metric is computed. Below a certain tolerance level, the statistical model that embeds the simulator may be deemed as not valid. The metric is named a highest posterior relative density defined as, for a new field data (not used in the computation of the predictive distribution):

$$\gamma(y_{new}) = 1 - \int \mathbb{I}\left(\left\{y : \frac{\pi(y|\mathbf{y}^e)}{q(y)} \geq \frac{\pi(y_{new}|\mathbf{y}^e)}{q(y_{new})}\right\}\right)\pi(y|\mathbf{y}^e)dy \tag{2.7}$$

where q is a reference distribution which can be taken as a uniform distribution, $\mathbb{I}(A)$ is the indicator function for the set A and $\pi(\cdot|\mathbf{y}^e)$ is the predictive distribution for a new observation conditioned to the considered field data. Note that this definition corresponds to the case of a cheap simulator. The definition can be extended to embed an emulator. This metric says to what extent an actual observation is plausible under the posterior predictive distribution. If this metric is smaller than a small threshold (e.g. 0.05, 0.01) for too many new field data, the prediction issued by the model are then suspect. Note that the predictive posterior distribution can be derived from a model with or without a discrepancy. In the former case, it can assess whether the simulator by itself captures sufficiently well the real process and in the latter case it can assess whether the discrepancy modeling is appropriate to compensate the bias of the simulator. The paper [132] and the section dedicated to model validation in [149] also deal with the cases where the outputs of the simulator are inaccessible to physical experiments. Only small experiments can be made and compared to some intermediate outcomes of the simulator.

2.2 Contributions

My contributions are presented below, grouped by methodological issues.

2.2.1 Sensitivity Analysis and Screening in Complex High Dimensional Simulators

The contributions in this section are mainly devoted to perform simple sensitivity analyses on complex simulators. The complexity is a result of the long computational time and the huge amount of outputs produced by the simulators. We proposed some tools to visualize the impacts of the inputs on the outputs at different levels (temporal, spatial and aggregated). An emerging difficulty stems from the mixture of various types of inputs. They vary in dimension: time-series, spatial map, scalar and in nature: quantitative or qualitative including the resolutions at which the simulator is run.

2.2.1.1 Analysis of Nitroscape Simulator

Presentation of NitroScape. In [JP4], we conduct an SA of NitroScape. NitroScape is a deterministic, spatially distributed and dynamic model describing Nitrogen (in its various chemical forms N_r) transfers and transformations in rural landscapes [60]. For each simulated day, it couples four modules characterizing farm management, biotransformations and transfers by the atmospheric and the hydrological pathways (see Figure 2.3(a)). It simulates the concentrations and fluxes, including the losses, of different forms of N_r (reduced forms (ammonia NH_3, ammonium NH_4^+), inorganic oxidized forms (nitrate NO_3^-, nitrogen oxides NO_x and nitrous oxide N_2O) and organic forms (manure, crop residues) within and between several landscape compartments: the atmosphere, the hydro-pedosphere (soil, water table, groundwater and streams) and the terrestrial agroecosystems (livestock buildings, croplands, grasslands and semi-natural areas).

Running NitroScape for SA. To run NitroScape we have to provide a spatial map describing the land use, climatic time-series giving amount of precipitations and temperatures and values for specific input parameters. The spatial map was chosen as a simplified theoretical landscape of 300 ha (Figure 2.3(b)) corresponding to an intensive rural area with succession of maize and wheat crops in a checkerboard distribution (125 ha each), pig farming buildings (two separate buildings, one ha each), and unmanaged grasslands (5 plots scattered, 48 ha in total). The topography was set as a linear slope with a gradient of 50 m. Meteorological data used for simulations were measured with a meteorological station located on the Kervidy-Naizin catchment (Brittany, 48°01'N, 2°83'O) between 2007 and 2011. It corresponds to humid climatic conditions and little temperature contrasts. 11 input factors were considered: the spatial (*i.e.* horizontal and vertical) resolution of the model (quantitative input factors A and B), the biophysical parameters which affect a priori the N_r fluxes in the agro-pedo-hydrosphere (quantitative input factors C to I) and two farm practices which mainly affect N_r fluxes and concentrations (qualitative and quantitative input factors J and K). Three levels were considered per factor. A fractional factorial design (FFD) [147] of size 243, that corresponds to 243 configurations combining the 11 input factors, was chosen of resolution 5. It then makes it possible to determine for each output the main effects and the pairwise interactions of input factors, without confounding effect between factors [21], in a model of analysis of variance (ANOVA). This design was also saturated since there was no residual degree of freedom to estimate the variance. From the simulations corresponding to the FFD configurations, the SA indices are computed in order to assess the effect of these 11 factors. The impact of the different land uses is quantified through the spatial analyses of the SA indices.

Simulations were performed at a daily time step and integrated over a five-year period, starting from January 1^{st}, 2007. The first two years of simulation were used for model initialization and the sensitivity analysis used the results provided by the last three years of simulation only. Daily outputs were sampled from the variables simulated at the catchment outlet and monthly outputs were sampled from results obtained at different locations within the landscape. Spatially-distributed outputs formed large sets of data that were difficult to handle with conventional statistical tools: each output was described by a matrix of 243 rows and up to more than 7.10^5 columns. Each row corresponded to each configuration of the FFD and each column corresponded to each output variable in each grid cell of the theoretical landscape. For instance, for the highest horizontal resolution (*i.e.* grid cells of size 12.5 m x 12.5 m each, Fig. 1b), the theoretical landscape included 19,600 grid cells, each characterized by the value of the 36 simulated monthly output variables, which resulted in 705,600 columns. For this reason, the output variables were spatially- or temporally-aggregated to produce different types of data sets: time series describing spatially-aggregated outputs were used to perform temporal sensitivity analysis, while maps of temporally-aggregated outputs were used for spatial sensitivity analysis. All output variables were also spatially- and temporally-aggregated to provide a synthetic view of the sensitivity of model outputs to input factors.

Sensitivity indices. For each configuration i of the FFD ($i = 1, \ldots, N$; $N = 243$), let Y_i be the outputs of interest ($Y_i = f(z_{i,1}, \ldots, z_{i,q})$; $q = 11$; factor number $j = 1, \ldots, q$ corresponding respectively to letters A to K). These outputs are scalar and have been obtained by spatially- or temporally-aggregation or by projection of the time series or the spatial map of a given output on one axis of a PCA. Note that we make a slight abuse of notation in this paragraph, by considering that f may correspond to different aggregations or projections of some outputs of the simulator. The notation $z_{i,j}$ stands for the input factor j of the configuration i of the FFD. The three different levels of each factor j are denoted by k ($k = 1, 2, 3$). An ANOVA model was adjusted to analyze main effects and second order interactions between factors:

$$Y_i = f(z_{i,1}, \ldots, z_{i,q}) = \mu + \sum_{j=1}^{q} \alpha^{(j)}_{z_{i,j}} + \sum_{1 \leq j < j' \leq q} \beta^{(j,j')}_{z_{i,j}, z_{i,j'}} + E_i$$

where $\alpha^{(j)}_{z_{i,j}}$ is the main effect of factor j on the output and $\beta^{(j,j')}_{z_{i,j}, z_{i,j'}}$ is the pairwise second order interactions between factors j and j' on the output, with $1 \leq j < j' \leq q$. These two effects were calculated by using the least squares method. The FFD being saturated, the residual terms E_i were all zero. The residual variance could not be therefore estimated. Since the NitroScape model is a deterministic model, the residual variance would have only corresponded to interactions of order higher than two. For a given output, the main effect of each factor j is:

$$mSI_j = \sum_{k=1}^{3} \# \mathscr{X}^{(k)}_j \cdot (\bar{Y}^{(k)}_j - \bar{Y})^2 \Big/ TSS$$

where $\bar{Y} = \frac{1}{n} Y_i$ is the overall average of Y_i's, $\mathscr{X}^{(k)}_j = \{1 \leq i \leq n : z_{i,j} = k\}$ are the sets of configurations i such that the factor j has level k, # denotes the cardinal of a set, $\bar{Y}^{(k)}_j = 1/\# \mathscr{X}^{(k)}_j \cdot \sum_{i \in \mathscr{X}^{(k)}_j} Y_i$ are the means for the levels k of factor j and $TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is the total sum of squares.

For each $1 \leq j < j' \leq q$, the pairwise interaction effects are given by:

$$SI_{j,j'} = \sum_{k,k'=1}^{3} \# \mathscr{X}^{(k,k')}_{j,j'} (\bar{Y}^{(k,k')}_{j,j'} - \bar{Y}^{(k)}_j - \bar{Y}^{(k')}_{j'} + \bar{Y})^2 \Big/ TSS$$

where $\mathscr{X}^{(k,k')}_{j,j'} = \{1 \leq i \leq n : z_{i,j} = k \text{ and } z_{i,j'} = k'\}$ are the sets of configurations i such that the factor j (resp. j') has level k (resp. k') and $\bar{Y}^{(k,k')}_{j,j'} = 1/\# \mathscr{X}^{(k,k')}_{j,j'} \cdot \sum_{i \in \mathscr{X}^{(k,k')}_{j,j'}} Y_i$. We also defined for each factor j an index summing up pairwise interaction effects involving this factor:

$$iSI_j = \sum_{j' : j' \neq j} SI_{j,j'},$$

an index describing the total (*i.e.* main and interaction) effect of factor j:

$$tSI_j = mSI + iSI_j,$$

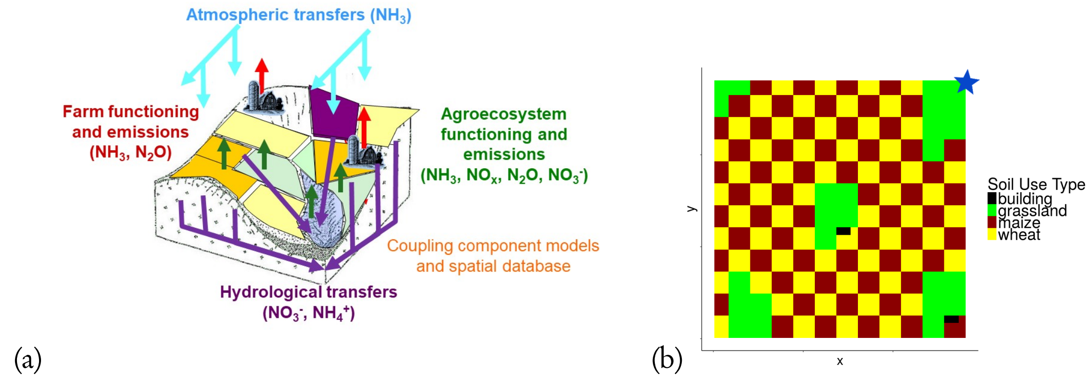

Atmospheric transfers (NH_3)
Farm functioning and emissions (NH_3, N_2O)
Agroecosystem functioning and emissions (NH_3, NO_x, N_2O, NO_3^-)
Coupling component models and spatial database
Hydrological transfers (NO_3^-, NH_4^+)
(a)
(b)
y
x
Soil Use Type
■ building
🟩 grassland
🟥 maize
🟨 wheat

Figure 2.3 – *Scheme of the NitroScape model (a). Land use and topography of the theoretical landscape (b), shown here for the highest spatial horizontal resolution of the model (grid cells of size 12.5 m x 12.5 m each). The blue star indicates the catchment outlet.*

and an index describing the sum of interactions between all factors:

$$i_{tot} = \sum_{1\leq j<j'\leq q} SI_{jj'} .$$

The FFD being saturated, the sum of the main effects of all factors (mSI_j) and of the ensemble of pairwise interactions (i_{tot}) added up to 100% of the total variance explored by the experimental design. Thus, i_{tot} was used as a direct measure of the variance that could not be attributed to any single factor.

Workflow and results. The workflow used to analyze the NitroScape model is described in Figure 2.4. Three levels of analyses are considered: a temporal analysis where the outputs are spatially-aggregated, a spatial analysis where the outputs are temporally-aggregated and a global analysis in which aggregation is both spatial and temporal.

For the global analysis, PCA was applied to the ensemble of sensitivity indices of the ensemble of temporally- and spatially-aggregated outputs, in order to better visualize the outputs that had similar responses to input factors and evaluate the relationship between the overall effects of the different factors. A hierarchical clustering and a PCA was applied to the data set **S** ($=(S_{ij})_{1\leq i\leq 243,\ 1\leq j\leq 66}$), in which each row corresponds to each of the 243 configurations of the FFD and each column corresponds to each of the 11 main sensitivity indices mSI_j and each of the 55 ($=\binom{11}{2}$) pairwise interaction indices iSI_j. Figure 2.5 displays the PCA with the 5 identified clusters of the 29 outputs of NitroScape. The clustering gather together outputs which are mostly impacted by the same inputs and the PCA helps to identify for these clusters which are these inputs. For instance, the light blue cluster gathers two outputs (NO_3^- concentration in groundwater and NH_4^+ concentration in soil) which are mainly sensitive to vertical resolution set when running the simulator. This figure also helps to identify the most important inputs and their joint influence. For instance, the two inputs C and D are often both influential or none of them is. When a simulator has many outputs, this plot allows the practitioner to have a quick glance at the main effects of the inputs.

The complete temporal analysis for NH_4^+ uptake by plants is provided in Figure 2.6. The projection on the three principal component of the time-series **S** =

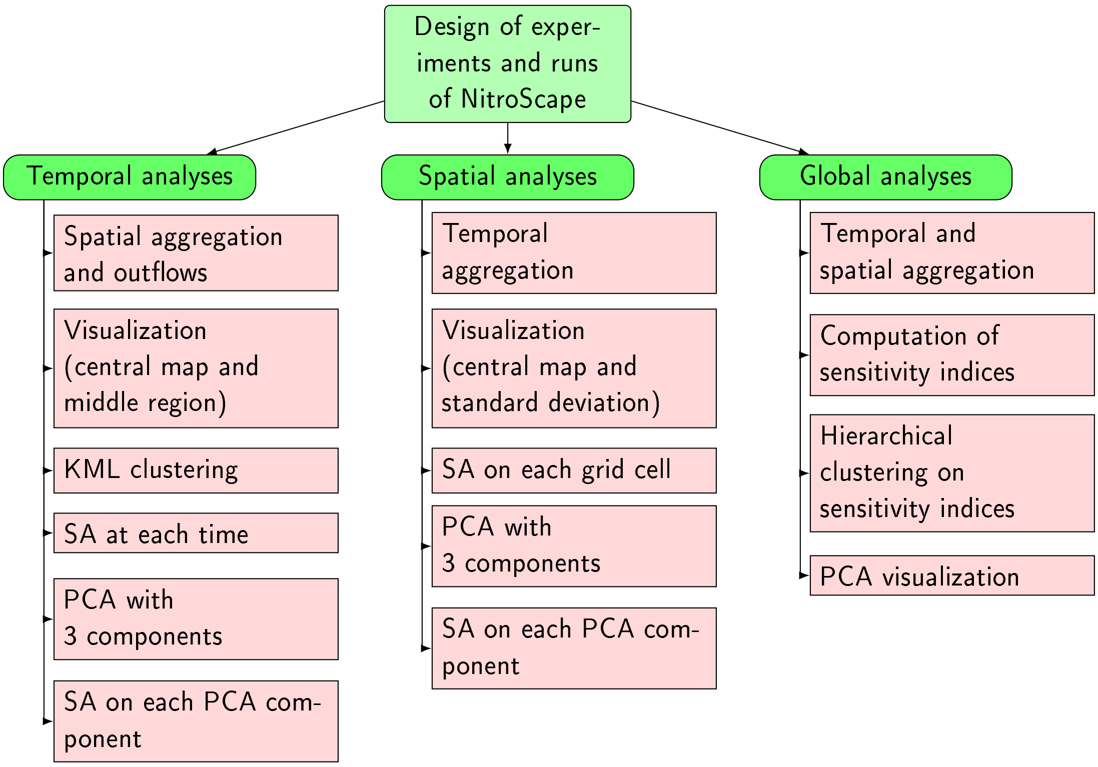

Design of experiments and runs of NitroScape
Temporal analyses
Spatial aggregation and outflows
Visualization (central map and middle region)
KML clustering
SA at each time
PCA with 3 components
SA on each PCA component
Spatial analyses
Temporal aggregation
Visualization (central map and standard deviation)
SA on each grid cell
PCA with 3 components
SA on each PCA component
Global analyses
Temporal and spatial aggregation
Computation of sensitivity indices
Hierarchical clustering on sensitivity indices
PCA visualization

Figure 2.4 – *Workflow of sensitivity analyses (see Subsection 2.2 for details). SA means Sensitivity analysis and PCA means Principal component analysis.*

$(Y_{it})_{1 \leq i \leq 243,\ 1 \leq t \leq 36}$ reduces data redundancy and identifies features linked to the model structure [103], such as seasonality, the first component being the average value and then the two others being related to seasonality. For each of these components, the sensitivity indices are computed. Clearly, the variation in the average value of NH_4^+ uptake (PC1) comes mainly from pairwise interactions involving soil surface porosity and fertilization type (factors F and J) while its seasonality (PC2) is more related to soil lateral transmissivity (factor C) and its interactions.

The complete spatial analysis for NH_4^+ uptake by plants is provided in Figure 2.7. The PCA is applied to the spatial maps $\mathbf{S} = (Y_{is})_{1 \leq i \leq 243,\ 1 \leq s \leq nc}$ where n_c is the total number of grid cells. PC1 describes roughly the spatial mean of FFD variance. PC1 was mostly sensitive to the main effect of soil surface porosity (factor F) and to its pairwise interactions. PC2 was strongly correlated with unmanaged grasslands downslope and less correlated with croplands and upslope areas.

Figures 2.6 and 2.7 represent two different aspects of the detailed sensitivity analysis of the accumulated NH_4^+ uptake by plants. The joint analyses of the spatially-aggregated and temporally-aggregated data sets made it possible to analyze the effects of the inputs on outputs from two complementary points of view and offered a more comprehensive visualization of the effects of the inputs.

2.2.1.2 Analysis of WALTer

In [JPC2], we study a simulator which simulates the individual development of plants and their interactions. An SA was conducted as in Section 2.2.1.1 on some scalar fea-

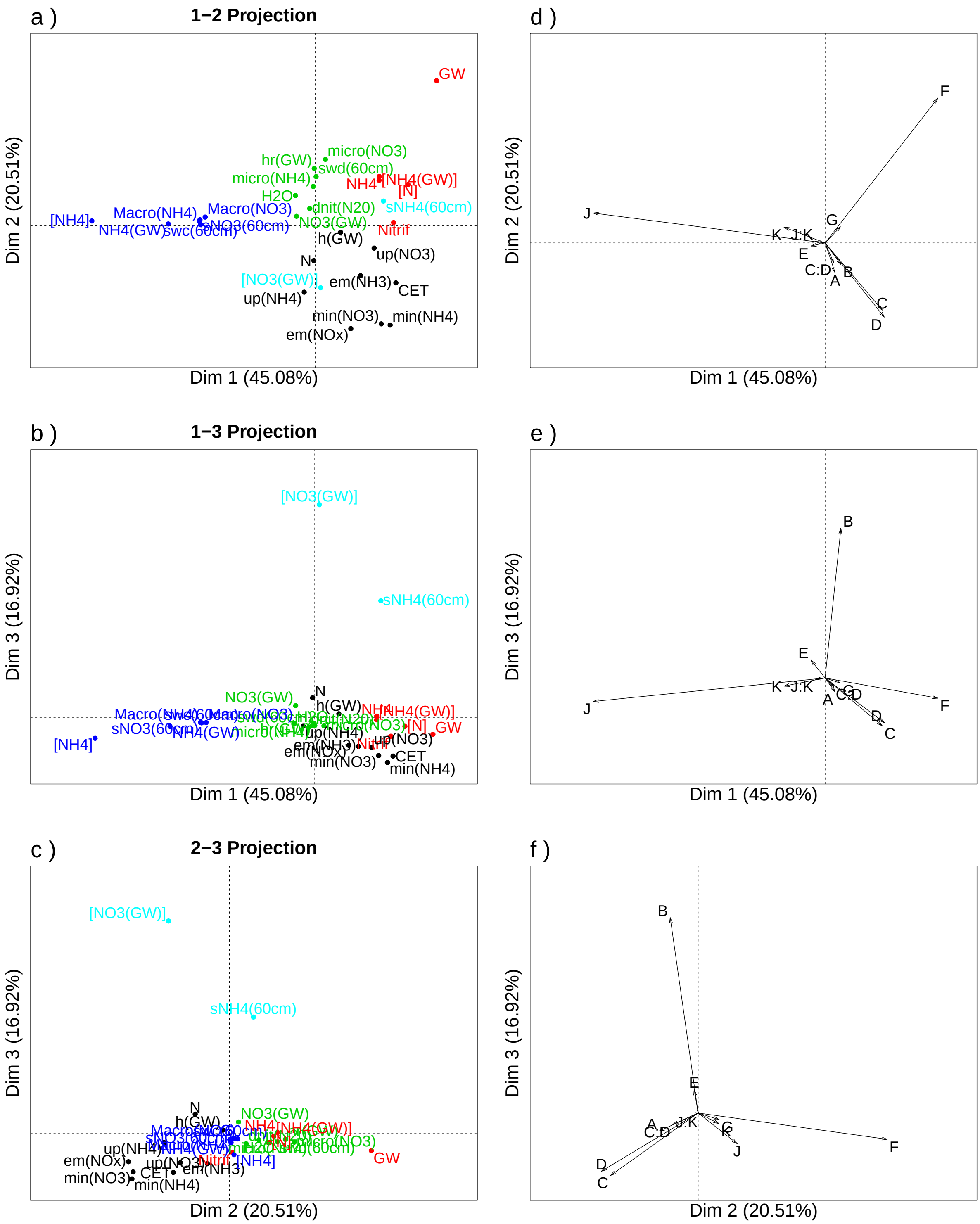


Figure 2.5 – *Principal component analysis and clustering of the results of the sensitivity indices resulting from the analysis of the 29 temporally- and spatially-aggregated outputs; (a,b,c) projections of the clusters of outputs onto the plane defined by two principal components; (d,e,f) projections of sensitivity indices of input factors onto the same planes.*

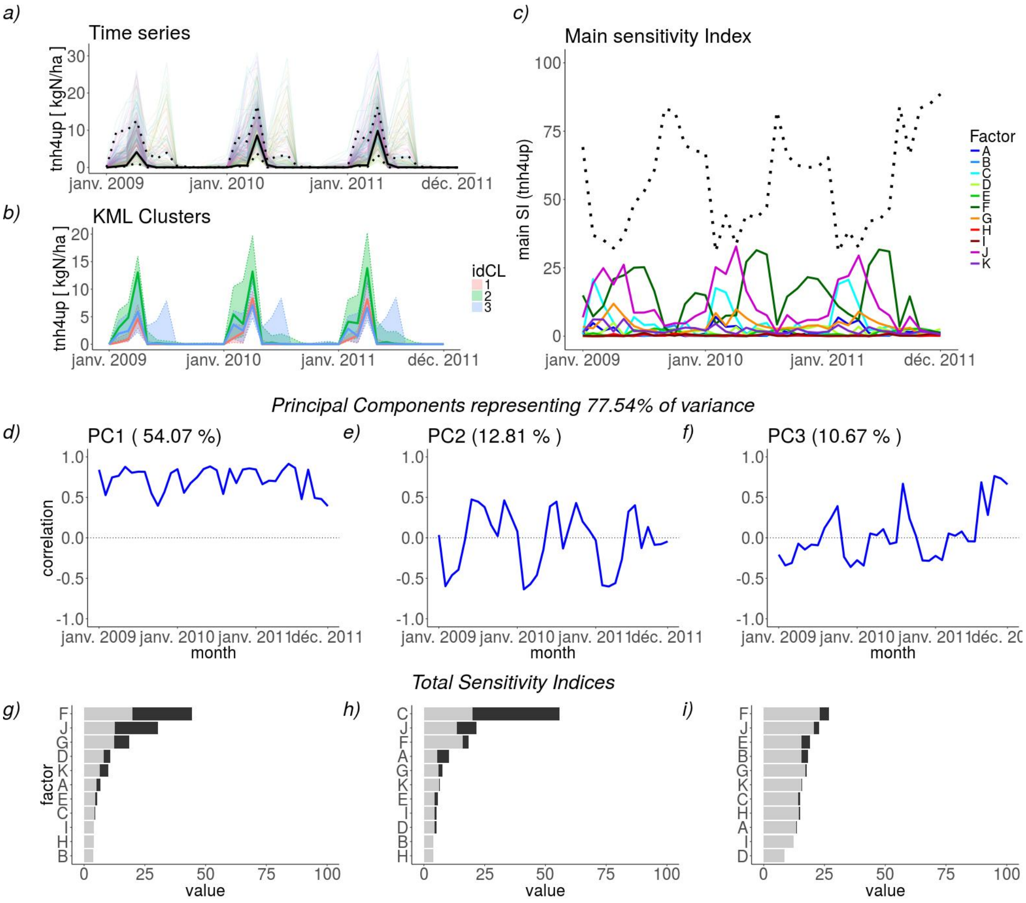


Figure 2.6 – *Temporal sensitivity analysis of NH_4^+ uptake by plants simulated for the whole landscape and averaged by area unit; (a) time series of each simulated configuration of the numerical experiment (colored lines), central time series (bold black line) and middle region (dashed black line); (b) time series of three clusters grouping most-similar curves, idCL is cluster label; (c) temporal main sensitivity indices of each factor (colored lines) and of the sum of interactions (dashed black line). Sensitivity analysis on each PC: (d,e,f) decomposition of the first three principal components (PC); (g,h,i) total sensitivity indices of each factor on each PC, split into main (black bars) and pairwise interaction (gray bars) effects.*

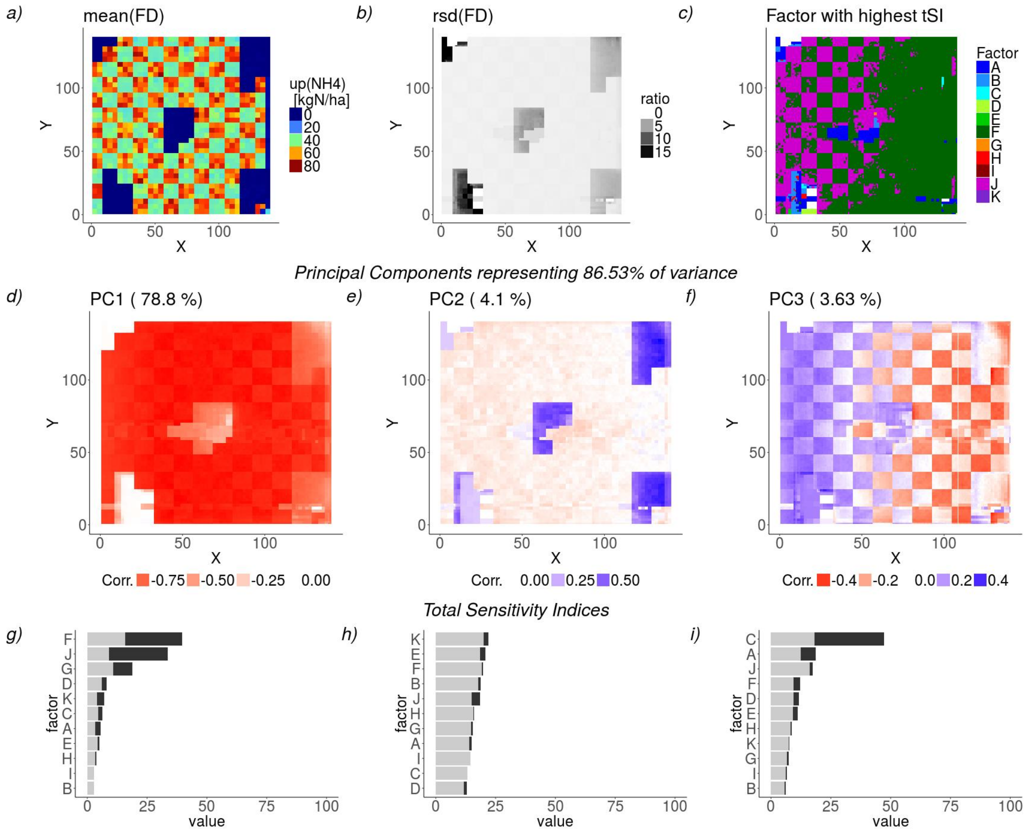


Figure 2.7 – *Spatial sensitivity analysis of NH_4^+ uptake by plants accumulated on the three-year period of interest in each grid cell of the landscape and averaged by area unit; (a) central map of averages over time within the fractional factorial design (FFD); (b) rsd: coefficient of variation between configurations of the FFD averaged over time; (c) map of the factors with the highest total sensitivity index (tSI) in each grid cell. Sensitivity analysis on principal components: (d,e,f) decomposition of the first three principal components; (g,h,i) total sensitivity indices of each factor on each PC, split into main (black bars) and interaction (gray bars) effects.*

tures extracted from the tillering dynamics. This first collaboration with plant biologists is an important source of new questions that will nourish the perspectives.

Presentation of WALTer. WALTer simulates the 3D development of the aerial architecture of winter wheat from sowing to flowering. The model is defined at plant scale, the crop being represented as a population of individual plants (Agent-Based Model). WALTer describes the plant architecture through a dynamic set of modules representing the plant components, their topology and geometry. A plant is composed of several axes (main stem and tillers from primary order to higher orders). WALTer includes both deterministic and adaptive processes. The development and extension of vegetative organs (blades, sheaths and internodes) follows descriptive rules. By contrast, tillering is described as a self-regulated process and modeled through two simple rules considering a critical Green Area Index (GAI) at which the emergence of tillers stops and a critical amount of light intercepted by each tiller under which tiller death is triggered. In order to simulate tiller regression, the interception of light by each tiller was computed by a radiative model (CARIBU) [35] applied to the 3D representations of plants, thus accounting for the competition for light among neighbor plants. The model is run with a daily time step and time is expressed as thermal time. For the sake of realism, some stochasticity is included in the model (final number of main stem leaves, probability of tiller emergence, duration before the plant emergence, plant and organ positions). This allows to model the effect of micro-environmental heterogeneity or the variability usually observed in plant development.

Sensitivity analysis. WALTer is based on more than 50 input parameters, so we selected a subset of input factors for sensitivity analysis. After discarding the parameters with values known with a good confidence from bibliography, we chose to investigate the effects of parameters with no/sparse experimental data and/or parameters directly impacting the tillering and GAI dynamics in WALTer formalism. Eventually, 8 inputs were selected for the SA. 6 inputs correspond to ecophysiological parameters: the critical GAI inducing cessation of tillering (GAIc), a threshold of intercepted light needed for the survival of a tiller (PARt), the protection duration between the death of two successive tillers of a plant (Δprot), the maximal length of the longest blade, the final number of leaf on the main stem and the range of the proximity GAI (dGAIp). Then two "environmental" parameters were also selected: sowing density, known to affect dramatically tillering, and Incident light, that defines the amount of photosynthetically active radiation incoming each day, on which is based tiller regression. The range of variation of each input parameter was set in a way to represent the largest space to explore with only three values. These values were chosen according to bibliographic and experimental data as well as exploratory simulations.

The two main outputs of interest of WALTer are (i) the tillering dynamics and (ii) the GAI dynamics, which provides information at the crop scale. We defined a set of scalar descriptors to summarize those dynamics in the most relevant way. The GAI dynamics was characterized by two scalar measures: the maximum value of the simulated GAI during the crop cycle (GAImax) and the date at which GAImax is reached (DGAImax). The tillering dynamics was characterized by four scalar measures: the maximum number of axes produced per plants ($\mathrm{N}_{\mathrm{axes}}^{\mathrm{max}}$), the duration of the tillering plateau ($\Delta_{\mathrm{plateau}}$), the number of ears produced per plant ($\mathrm{N}_{\mathrm{ears}}$) and the rate of tiller

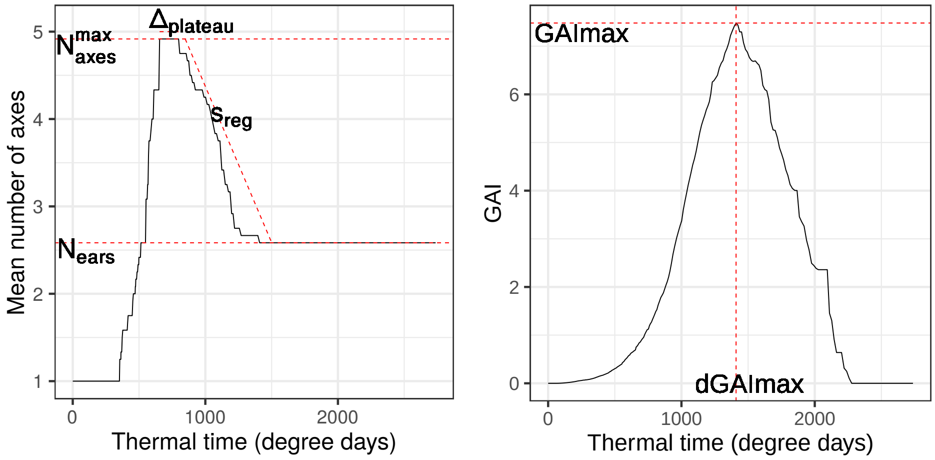


Figure 2.8 – *Tillering dynamic (on the left-hand-side) and GAI dynamic (on the right-hand-side) issued by WALTer for a specific set of inputs. The number of axes and the GAI are averaged over all the simulated plants. The four quantities of interests derived from the tillering dynamic are represented on the LHS: N^{max}_{axes}, $\Delta_{plateau}$, N_{ears} and s_{reg} and the two ones derived from the GAI dynamic are represented on the RHS: GAImax and DGAImax.*

regression (s_{reg}). See Figure 2.8 for an illustration of these quantities of interest on a run of WALTer.

These outputs are produced for every plant in a simulation, then they are averaged on the central plants in the simulated parcel to discard border effects. The number of plants was set to 200 on the basis of first simulations in order to limit the effect of the stochasticity of the simulator.

The SA relies on the same methods as in Section 2.2.1.1 with the same FFD of resolution 5 with 243 configurations. Since the number of inputs is 8, the FFD is not saturated and the residuals can be estimated. In this case, it corresponds to interactions of order larger than 2 and of the stochasticity embedded in WALTer.

The main outcomes of the SA is the predominant importance of the critical GAI and the threshold (PARt).

Calibration and validation. In order to test the ability of WALTer to predict the GAI and the tillering dynamics, some data with different sowing densities were used [48]. The dynamics corresponding to an average density (200 plants per squared meter) was used to calibrate the most influential parameters of WALTer. This calibration was done quite heuristically by minimizing a mean square error between the field data and the outputs of WALTer. With the notation previously introduced, it consists in estimating $\hat{\theta}$ such that

$$\hat{\theta} = \arg\min_{\theta} \sum_t (y^e_t - f(x = 200, \theta)_t)^2 .$$

In this case the input variable x is the density and the output is indexed by time since WALTer gives tillering dynamics. Then, the validation consists in running WALTer with the calibrated parameters (i.e. evaluation $f(x, \hat{\theta})$ for the other $x = x^e$ in the dataset) for the other densities and to compare with the field data.

The comparison led to satisfying results but some discrepancies appeared. With respect to some features WALTer had a correct behavior, e.g. the duration of the plateau of tillering increased with increasing density as in the dataset even if the duration did not exactly match the ones in the dataset.

Properly accounting for the specific stochastic nature of the simulator in a well grounded statistical model is one of the perspectives coming from this work.

2.2.1.3 Analysis of a Crop Circulation Network

In [JP12], we studied a dynamic extinction colonization model (also called contact process) which is a wide subject in epidemiology and in metapopulation theory. Contacts are usually assumed to be possible only through a network of connected patches. This network accounts for a spatial landscape or a social organisation of interactions. A major issue is to assess the influence of the network in the dynamic model. The network has to be reduced to simple topological descriptors in order to assess its impact on the dynamic process. This question is related to the work on network presented in Chapter 3. It is specifically dealt with in Section 3.2. We can however emphasize here that the stochasticity matters in this simulator. Indeed, the colonization or infection events are generally assumed to be stochastic and the network has a limited size which makes the stochasticity important. Therefore, some scalar features are extracted from replicates. The impact of the topological descriptors of the networks on these features was then assessed by an SA derived from an ANOVA model.

2.2.2 Inverse Problems and Calibration

In the papers [JP6, JP2, JP7], we considered inverse problems in the sense that we aimed to estimate from observations modeled as outputs of the simulator, some of its input parameters. This task is a calibration task in [JP6, JP2] as presented in Section 2.1.2 while there is an extra layer in the hierarchical model considered in [JP7] since the input parameters are drawn in a probability distribution. The latter is a random effect model with different realizations of the parameters corresponding to different observations. In [JP6, JP2] the inference is Bayesian while it relies on a stochastic version of the EM algorithm in [JP7].

2.2.2.1 Calibration

Statistical models and corresponding likelihood. In [JP2], we focused on variations around Equation (2.5). We considered that the simulator needs either to be emulated or not and that a discrepancy function is either added or not. This results in four different statistical models. We denote by $\mathscr{M}_1$ the model with no emulation and no discrepancy, $\mathscr{M}_2$ the model with emulation but no discrepancy, $\mathscr{M}_3$ the model with no emulation but discrepancy and $\mathscr{M}_4$ the model with both emulation and discrepancy. When the discrepancy is taken into account in the model, we assume a zero mean $\delta \sim \mathscr{GP}(0, C_\delta(\cdot,\cdot))$

To perform the calibration we derive the four likelihoods. For Models $\mathscr{M}_1$ and $\mathscr{M}_3$ the likelihoods read as:

$$\ell(\theta, \sigma_\epsilon^2, \psi_\delta, \sigma_\delta^2; \mathbf{y}^e) = \frac{1}{(2\pi)^{n_e/2}|\mathbf{V}_e|^{1/2}} \exp\left\{ -\frac{1}{2}\left(\mathbf{y}^e - \mathbf{m}_e\right)^T \mathbf{V}_e^{-1}\left(\mathbf{y}^e - \mathbf{m}_e\right)\right\}. \quad (2.8)$$

where $\mathbf{m}_e = \mathbf{m}_e(D^e, \theta) = (f(\mathbf{x}_i^e, \theta))_{1 \le i \le n_e}$, $\mathbf{V}_e = \sigma_\epsilon^2 I_{n_e}$ in $\mathscr{M}_1$ and $\mathbf{V}_e = \mathbf{V}_e(D^e) = \sigma_\epsilon^2 I_{n_e} + \sigma_\delta^2 (C_\delta(\mathbf{x}_i^e, \mathbf{x}_j^e))_{1 \le i,j \le n_e}$ in $\mathscr{M}_3$. Note that C_δ does depend on the parameters

ψ_δ. The parameters $\psi_\delta, \sigma_\delta^2$ are in gray in the equation above since they intervene only for $\mathcal{M}_3$.

For Models $\mathcal{M}_2$ and $\mathcal{M}_4$, we rely on modularization i.e. the likelihood of the field data $\mathbf{y}^e$ conditioned to $f(D)$ where the MLE of the parameters of the GPE have been plugged in is under consideration for calibration:

$$\ell^C(\theta, \sigma_\epsilon^2, \psi_\delta, \sigma_\delta^2; \mathbf{y}^e | f(D)) \propto |\mathbf{V}_{e|f(D)}|^{-1/2} \exp\Big\{ -\frac{1}{2}(\mathbf{y}^e - \mathbf{m}_{e|f(D)})^T \mathbf{V}_{e|f(D)}^{-1} (\mathbf{y}^e - \mathbf{m}_{e|f(D)}) \Big\}. \tag{2.9}$$

with $\mathbf{m}_{e|f(D)} = \mathbf{m}_{e|f(D)}(D^e, \theta) = (m_D(\mathbf{x}_i^e, \theta))_{1 \le i \le n_e}$ and $\mathbf{V}_{e|f(D)} = \mathbf{V}_{e|f(D)}(D^e, \theta)$ with

$$\begin{array}{rcl} \mathbf{V}_{e|f(D)} & = & \sigma_\epsilon^2 I_{n_e} + (C_D((\mathbf{x}_i^e, \theta), (\mathbf{x}_j^e, \theta)))_{1 \le i,j \le n_e} \text{ in } \mathcal{M}_2, \\ \mathbf{V}_{e|f(D)} & = & \sigma_\epsilon^2 I_{n_e} + (C_D((\mathbf{x}_i^e, \theta), (\mathbf{x}_j^e, \theta)))_{1 \le i,j \le n_e} + \sigma_\delta^2 (C_\delta(\mathbf{x}_i^e, \mathbf{x}_j^e))_{1 \le i,j \le n_e} \text{ in } \mathcal{M}_4. \end{array}$$

The function m_D and the covariance kernel C_D are given in Equation (2.3) with the parameters of the GPE fixed to plugged in estimates. The parameters $\psi_\delta, \sigma_\delta^2$ are in gray in the equation above since they intervene only for $\mathcal{M}_4$.

In [JP6], $\mathcal{M}_2$ is posited. Once prior distributions are set on the parameters, their posterior distributions can be sampled by combining these prior distribution and the likelihood in a Metropolis-Hastings algorithm [83].

Design of numerical experiments for calibration. In the previous paragraph, the likelihoods which integrate a GPE (in $\mathcal{M}_2$ and $\mathcal{M}_4$) can be obtained from any DoNE D and the corresponding evaluations $f(D)$. However, some designs may lead to a better calibration than others. If there is no discrepancy, the gold standard when using a GPE is to obtain a posterior distribution $\pi_2(\theta|\mathbf{y}^e, f(D))$ as close as possible to the posterior distribution $\pi_1(\theta|\mathbf{y}^e)$. The subscript 1 and 2 in $\pi_{1,2}$ refer respectively to Models $\mathcal{M}_1$ and $\mathcal{M}_2$. Under the assumption that σ_ϵ^2 is known, the posterior distributions $\pi_1(\theta|\mathbf{y}^e)$ and $\pi_2(\theta|\mathbf{y}^e, f(D))$ are proportional to likelihood coming from respectively Equations (2.8) and (2.9) times a prior distribution $\pi(\theta)$:

$$\begin{array}{rcl} \pi_1(\theta|\mathbf{y}^e) & \propto & \ell(\theta; \mathbf{y}^e) \cdot \pi(\theta) \\ \pi_2(\theta|\mathbf{y}^e, f(D)) & \propto & \ell^C(\theta; \mathbf{y}^e | f(D)) \cdot \pi(\theta). \end{array}$$

The Kullback-Leibler (KL) divergence shows interesting theoretical properties to measure how far a probability distribution is from a reference one [42]. It reads as

$$\mathrm{KL}\Big(\pi_1(\theta|\mathbf{y}^e) || \pi_2(\theta|\mathbf{y}^e, f(D))\Big) = \int_\Theta \pi_1(\theta|\mathbf{y}^e) \Big(\log(\pi_1(\theta|\mathbf{y}^e)) - \log(\pi_2(\theta|\mathbf{y}^e, f(D)) \Big) d\theta. \tag{2.10}$$

By using results of approximation theory, we can prove the proposition below.

Proposition 1. *Under the following assumptions:*

A1 *$\pi(\theta)$ has a bounded support Θ,*

A2 *the simulator output $f(\mathbf{x}, \theta)$ is uniformly bounded on $\mathcal{X} \times \Theta$,* *A3 the correlation function (kernel) is a classical radial basis function [151] i.e. there exists a function k such that* $C((\mathbf{x}',\theta'),(\mathbf{x},\theta)) = k(||(\mathbf{x}',\theta')-(\mathbf{x},\theta)||)$ *where* $||\cdot||$ *can be chosen as the Euclidean norm,*

A4 the function f lies in the Reproducing Kernel Hilbert Space associated with the kernel defining the correlation function,

A5 the covering distances associated with the sequence of DoNE $(D_M)_M$*:*

$$h_{D_M} = \max_{(\mathbf{x},\theta)\in\mathscr{X}\times\Theta}\ \min_{(\mathbf{x}_i,\theta_i)\in D_M} ||(\mathbf{x},\theta)-(\mathbf{x}_i,\theta_i)|| \xrightarrow[M\to\infty]{} 0.$$

then, we have:

$$\lim_{M\to\infty} KL\big(\pi_1(\theta|\mathbf{y}^e)||\pi_2(\theta|\mathbf{y}^e, f(D_M))\big) = 0. \tag{2.11}$$

The question is how to choose a limited number of points for D in order to make the KL divergence from Equation (2.10) the smallest as possible. The heuristic defended is this work is that the GPE should be close to the true simulator f especially for input configurations $(\mathbf{x}^e_i,\theta)_{1\le i\le n_e}$ with θ corresponding to a high value of the posterior distribution $\pi_1(\theta|\mathbf{y}^e)$.

Such a design D can actually be obtained as a natural by-product of a sequential and global maximization procedure for searching $\max_\theta \pi_1(\theta|\mathbf{y}^e)$. This procedure allocates the budget of simulation between locations where $\pi_1(\theta|\mathbf{y}^e)$ is high with respect to the θ coordinate and ones where exploration is needed. By doing so, the code is likely to have been run over values of θ which lie mainly in all the regions where $\pi_1(\theta|\mathbf{y}^e)$ is high. By using the log scale and neglecting terms which do not depend on θ, the maximization problem is equivalent to solving

$$\max_\theta -||\mathbf{y}^e - \mathbf{m}_e||^2/2\sigma_\epsilon^2 + \log(\pi(\theta)). \tag{2.12}$$

When little knowledge is available on the value of θ, either a uniform prior (if both a lower and an upper bound are provided) or a locally uniform prior is usually specified for θ [22].

In such cases, when there is substantial information in the data, the regions of high probability for $\pi(\theta|\mathbf{y}^e)$ are where $SS(\theta) = ||\mathbf{y}^e - \mathbf{m}_e||^2$ is small. In the following, we present our algorithms for constructing D in these cases. They are therefore based on the sequential minimization of $SS(\theta)$. Hence, the construction of the design D will be independent on the value of σ_ϵ^2. When the likelihood on θ is flat or if an informative prior is available, the construction of the design can be based on the optimization problem (2.12) which takes into account the prior at no additional cost. In this latter case, the construction of the design will depend on the value of σ_ϵ^2 since it balances the weight given to the sum of squares and the one given to the prior.

The Expected Improvement criterion was introduced [89] to find the global extremum of an expensive simulator and its location. We resort to this EI criterion for the sum of squares of the residuals function $SS(\theta)$:

$$EI_k(\theta) = \mathbb{E}\Big[\big(s_k - SS_k(\theta)\big)\mathbb{I}_{SS_k(\theta)\le s_k}\Big] \in [0, s_k], \tag{2.13}$$

where

- $s_k := \min\{SS(\theta_1),\cdots,SS(\theta_{k-1}),SS(\theta_k)\}$ and $SS(\cdot)$ denotes the sum of squares computed from actual runs of the computer code f,

- $SS_k(\theta)$ denotes the sum of squares of the residuals where $f(\mathbf{x},\theta)$ is replaced with the random vector $F^{D_k}(\theta) = \left(F^{D_k}(\mathbf{x}_1^e,\theta),\cdots,F^{D_k}(\mathbf{x}_n^e,\theta)\right)$, the distribution of which is given by the GPE conditional to $f(D_k)$:
$$SS_k(\theta) = ||\mathbf{y}^e - F^{D_k}(\theta)||^2 .$$

Note that the subscript k refers here to the current iteration of the algorithm. $SS_k(.)$ is thus a random process and its distribution inherits from the current GPE. At step k, once we have found θ_{k+1} maximizing EI_k, we could add n_e new simulations $\{f(\mathbf{x}_i^e,\theta_{k+1})\}_{1\leq i\leq n_e}$ to compute $SS(\theta_{k+1})$. This may be infeasible if n_e is not small. Thus we propose an approximate version (see[JP6] for the exposition of the full computational algorithm) where a location $\mathbf{x}^*$ is chosen among the field locations ($\mathbf{X}^e = \{\mathbf{x}_1^e,\ldots,\mathbf{x}_{n_e}^e\}$) so that only one run of the simulator f is run. Thus $SS(\theta_k)$ cannot be computed exactly. We use instead its expectation under the distribution of the current GPE. The location $\mathbf{x}^*$ is chosen as the optimum of a criterion Crit which can be chosen as:

$$\mathrm{Crit}(\mathbf{x}_i^e,\theta_{k+1}) = \mathrm{Var}[F^{D_k}(\mathbf{x}_i^e,\theta_{k+1})]. \tag{2.14}$$

This criterion aims to reduce the variance of the GPE where it is larger. Yet, a better way might perhaps consist in aiming for a reduction of the GPE uncertainty at an input location $(\mathbf{x}^*,\theta_{k+1})$ where the code $f(\mathbf{x}^*,\theta)$ is highly variable with respect to θ, meaning that $\mathbf{x}^*$ is influential for calibration. We thus introduce a second criterion which does a trade-off between the calibration goal and (2.14). A normalized version of it is written as

$$\mathrm{Crit}(\mathbf{x}_i^e,\theta_{k+1}) = \frac{\mathrm{Var}\left(F^{D_k}(\mathbf{x}_i^e,\theta_{k+1})\right)}{\max\limits_{i=1,\cdots,n}\mathrm{Var}\left(F^{D_k}(\mathbf{x}_i^e,\theta_{k+1})\right)} \times \frac{\mathrm{Var}_\theta[f(\mathbf{x}_i^e,\theta)]}{\max\limits_{i=1,\cdots,n}\mathrm{Var}_\theta[f(\mathbf{x}_i^e,\theta)]}, \tag{2.15}$$

where $\mathrm{Var}[y_\theta(\mathbf{x}_i^e)]$ is taken with respect to $\pi(\theta)$. In practice, we need to use an approximation of (2.15) that is based on the mean of F^{D_k}. The whole sequential design procedure is described in Algorithm 1.

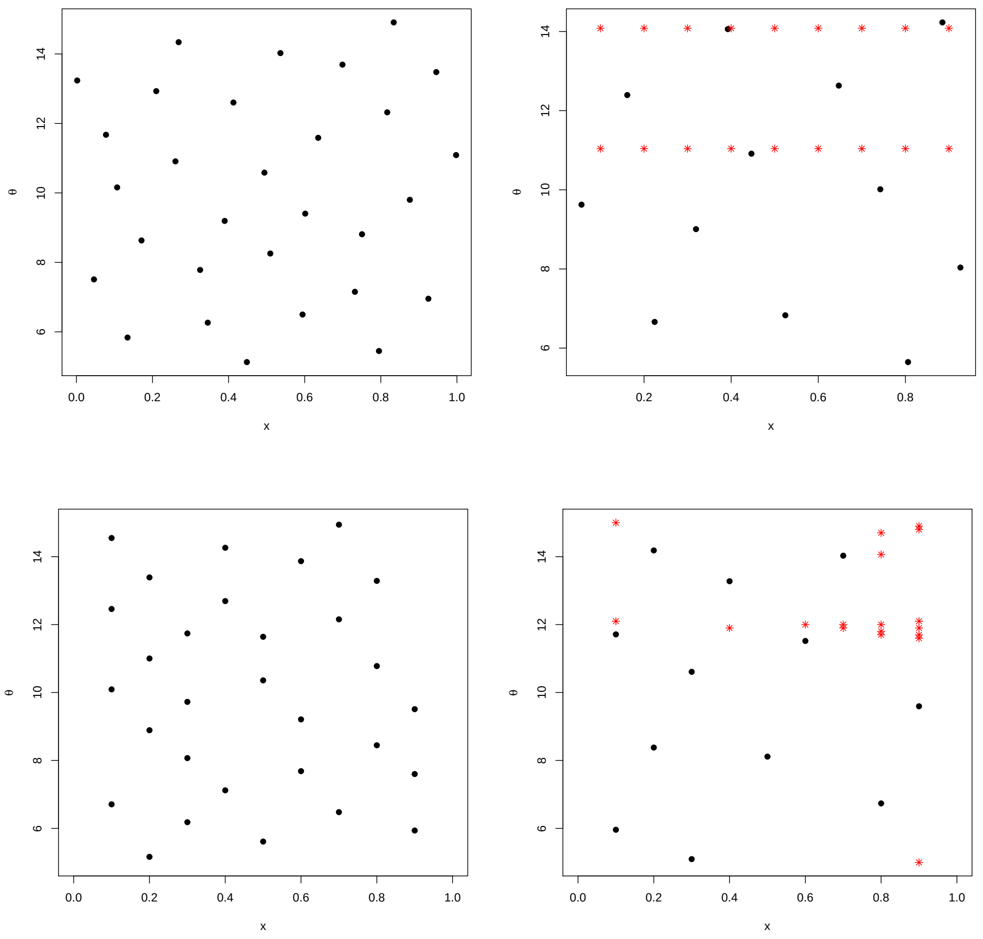
Remark. Note that D_0 in Algorithm 1 may be chosen as being space-filling in $\mathscr{X}\times\Theta$ or the coordinates in $\mathscr{X}$ may be restricted to be in the subset $\mathbf{X}^e = \{\mathbf{x}_1^e,\ldots,\mathbf{x}_{n_e}^e\}$ to correspond to actual field experiments.

We compare the proposed sequential designs with classical space-filling designs (restricted or not to the domain $\mathbf{X}^e$ for the input variables). We consider the toy simulator:

$$f:(x,\theta)\in[0,1]\times[5,15]\longrightarrow(6x-2)^2\times\sin(\theta x-4), \tag{2.16}$$

with the field design $\mathbf{X}^e = \{.1,.2\ldots,.9\}$. The field observation $\mathbf{y}^e$ were simulated with $\theta = 12$. Figure 2.9 provides a comparison between space-filling designs and the sequential designs we propose. In space filling design, the exploration is uniform over the dimension in x and in θ. With the sequential design, it is observed that the exploration over θ concentrates around the true value $\theta = 12$ in the sequential step. Moreover, the exploration is reinforced for some values of x which makes the simulator more sensitive to θ. Note that the sequential strategy with the full batch of points $\{(\mathbf{x}_i^e,\theta_{k+1})\}_{1\leq i\leq n_e}$ is clearly not viable since it requests too many evaluations of f at each step of the algorithm. In [JP6], we show on artificial examples that the posterior distribution $\pi_2(\theta|\mathbf{y}^e,f(D)$ is closer to $\pi_1(\theta|\mathbf{y}^e)$ when D is a sequential design generated by Algorithm 1 than when D is a space-filling design.

Figure 2.9 – *DoNE for simulator 2.16 with $M = 30$ call to f. Upper left: maximin-LHS design. Upper right: Sequential design with* 12 *starting points sampled as a maximin design restricted to* $\mathbf{X}^e$ *and* 18 *sequential points added in 2 batches according to the EI criterion (Algo 1 in [JP6]). Bottom left: maximin design with restricted to* $\mathbf{X}^e$*. Bottom right: Sequential design with* 12 *starting points sampled as a maximin design restricted to* $\mathbf{X}^e$ *and* 18 *sequential points added one-at-a-time according to the EI criterion with criterion given in Eq* (2.15) *(Algo 1). The black dots are the initial design. The red stars are the new runs selected from the EI criterion.*



```
Algorithm 1: Sequential Design Adapted to Calibration
  Initialization
    • Choose an initial numerical design $D_0 \subset \mathscr{X} \times \Theta$ of size $M_0$.
    • Run the code over $D_0$, then obtain an initial GPE based on $f(D_0)$.
    • Compute $\hat{\theta}_1$ as the posterior mean $\mathbb{E}[\theta|\mathbf{y}^e, f(D_0)]$.
    • $D_1 = D_0 \cup \{(x_i^e, \hat{\theta}_1)\}$ where
          $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathbf{X}^e} \mathrm{Crit}(\mathbf{x}, \hat{\theta}_1)$.
    • Compute $s_1 := \mathbb{E}(SS_0(\hat{\theta}_1))$.
  From $k = 1$, repeat the following steps as long as $M_0 + 1 + k \le M$.
  Step 1 Find of $\theta^\star_{k+1} = \arg\max_\theta EI_k(\theta)$.
  Step 2 $D_{k+1} = D_k \cup \{(\mathbf{x}^*, {\theta^*}_{k+1})\}$ where
          $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathbf{X}^e} \mathrm{Crit}(\mathbf{x}, {\theta^*}_{k+1})$.
  Step 3 Run the code for the new location $(\mathbf{x}^*, {\theta^*}_{k+1})$.
  Step 4 Update the GPE distribution based on $f(D_{k+1})$.
  Step 5 Compute $s_{k+1} := \min\{\mathbb{E}[SS_k(\hat{\theta}_1)], \cdots, \mathbb{E}[SS_k({\theta^*}_k)], \mathbb{E}[SS_k({\theta^*}_{k+1})]\}$.
```

Table 2.1 – *Comparison of the RMSEs and coverage rates in prediction of 100 test-sets on three randomly selected days where $\mathscr{M}_2'$ and $\mathscr{M}_4'$ are the models based on the Gaussian process established after the sequential design*

	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$	$\mathscr{M}_4$	$\mathscr{M}_2'$	$\mathscr{M}_4'$
coverage rate at 90% (in %)	91	44	85	42	71	68
RMSE of power (W)	5.103	21.79	4.56	18.78	10.94	9.29

Results on a PV simulator. In [JP2], we were provided by EDF with a fast simulator of a photovoltaic (PV) power plant. The simulator $f(\mathbf{x}, \theta)$ outputs the power produced depending on a vector of meteorological conditions $\mathbf{x} \in \mathbb{R}^4$ measured in field data and on parameters $\theta \in \mathbb{R}^3$ describing the characteristics of the PV panels. Data from a test stand of 12 panels are available over 2 months and instantaneous power was collected every $10s$. To limit the size of the data set, we averaged the power per hour and we remove the data where the production is zero. It results in 1019 observations. Since the simulator is fast, Model $\mathscr{M}_1$ or $\mathscr{M}_3$ should be used for calibration and prediction. In order to assess the effect of the additional layer of uncertainty due to the use of an emulator, we also calibrated the simulator under Models $\mathscr{M}_2$ and $\mathscr{M}_4$. The accuracy of predictions under the different models were compared by cross-validation. Three days of data were successively removed (51 observations) from the learning dataset and were used as a test dataset. Table 2.1 provides the coverage rates and the RMSEs (Relative Mean Square Errors) under the four models. For Models $\mathscr{M}_2$ and $\mathscr{M}_4$, we use either a regular space-filling DoNE or a sequential DoNE that was generated with Algorithm 1. This shows on a real application that the sequential design significantly improves both the RMSEs and the coverage rate. This improvement is similar for Models $\mathscr{M}_2$ and $\mathscr{M}_4$ even though the sequential procedure in the algorithm was originally designed under the assumption of zero discrepancy. Note that the whole process from calibration through prediction to cross-validation was implemented in the `R`-package `CaliCo` [P7].

2.2.2.2 Mixed non Linear Models

The framework of the mixed model is close to the calibration context but the difference lies in the fact that the field data $\mathbf{y}^e$ are not all generated with the same value of the parameter. More precisely, we assume that the field data are $\mathbf{y}^e = (y^e_{ij})_{1\leq i\leq n_e, 1\leq j\leq n_i}$ where the vectors $\mathbf{y}^e_i = (y^e_{ij})_{1\leq j\leq n_i}$ are independent. The dependence within the vector $\mathbf{y}^e_i$ is the result of the generation of the y_{ij}'s with the same parameter value. More precisely, the mixed model reads as: for $i = 1,\dots,n_e$, $j = 1,\dots,n_i$:

$$\begin{array}{lll} y^e_{ij} & = & f(\mathbf{x}^e_{ij},\psi_i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim_{iid} \mathscr{N}(0,\sigma^2_\epsilon) \\ \psi_i & \sim_{iid} & \mathscr{N}(\mu,\Omega) \end{array} . \tag{2.17}$$

In this model we denote by $\theta = (\mu,\Omega)$ the parameters of the distribution of the latent parameters ψ_i's. Moreover, we do not assume discrepancy in this model. A classical framework for this mixed model is longitudinal data where the vectors $\mathbf{y}^e_i$ are observations in time for an individual. In this case, the input variables $\mathbf{x}^e_i$ contain the time of the observation. The latent parameters ψ_i account for the specificity of an individual and are called individual parameters while the parameters θ are called population parameters. In this model, the goal is to infer jointly $(\theta,\sigma^2_\epsilon)$. The likelihood corresponding to the mixed model (2.17) is

$$\begin{array}{ll} & \ell(\theta,\sigma^2_\epsilon;\mathbf{y}^e) = \int \ell(\theta,\sigma^2_\epsilon,\psi;\mathbf{y}^e)\,d\psi = \prod_{i=1}^{n_e} \int \pi(\mathbf{y}^e_i|\psi_i,\sigma^2_\epsilon)\pi(\psi_i|\theta)\,d\psi_i \\ = & \prod_{i=1}^{n_e} \int \left\{ \pi(\psi_i|\theta)\frac{1}{(2\pi\sigma^2_\epsilon)^{n_i/2}} \right. \\ & \left. \times \exp\left(-\frac{1}{2}(\mathbf{y}^e_i - f(\mathbf{x}^e_i,\psi_i))^t(\sigma^2_\epsilon I_{n_i})^{-1}(\mathbf{y}^e_i - f(\mathbf{x}^e_i,\psi_i))\right)d\psi_i \right\} \end{array} , \tag{2.18}$$

where the distributions $\pi(\psi_i|\theta)$ and $\pi(\mathbf{y}^e_i|\psi_i,\sigma^2_\epsilon)$ are given by Equation (2.17).

A classical solution to deal with a latent variable model is to resort to an EM (Expectation-Maximization algorithm) [55]. If the function f is not linear in ψ, the expectation step is not explicit. Then, a standard solution is to use stochastic version of the EM algorithm such as SAEM (Stochastic Approximation Expectation Maximization) [54] or SEM (Stochastic Expectation Algorithm) [29]. An alternative solution is to rely on Bayesian inference where the latent variables are re-simulated within a Gibbs algorithm. The simulator f may be the solution of an ODE or PDE. If the solution is not explicit, the inference procedures are combined with numerical resolution schemes [57]. This resolution scheme may be costly and then the GP emulation can be integrated in the inference procedure as was done within an SEM algorithm [JP18] or in a Bayesian inference [66]. In [JP7], we coupled the GPE with an SAEM algorithm. To simulate the latent parameters ψ, we resort to an MCMC algorithm as suggested in [99]. This step needs many calls to f which is unfeasible if f is costly. Thus, the GPE can be used in the likelihood (2.18) to alleviate the computational burdensome. A modular approach is used where the posterior distribution of the GP is plugged in the likelihood:

$$
\begin{aligned}
\ell_D(\theta,\sigma_\epsilon^2;\mathbf{y}^e) \;=\; \int \Bigg\{ & \pi(\psi_i|\theta)\, \tfrac{1}{(2\pi)^{n_{tot}/2}|\mathbf{V}_{e|f(D)}|^{1/2}} \\
& \exp\Big(-\tfrac{1}{2}(\mathbf{y}^e - \mathbf{m}_{e|f(D)})^t (\mathbf{V}_{e|f(D)})^{-1} (\mathbf{y}^e - \mathbf{m}_{e|f(D)}) \Big) d\psi \Bigg\},
\end{aligned}
\tag{2.19}
$$

where $n_{tot} = \sum_{i=1}^{n_e} n_i$, $\mathbf{m}_{e|f(D)} = (m_D(\mathbf{x}_{ij}^e,\psi_i))_{1\le i\le n_e, 1\le j\le n_i}$ and $\mathbf{V}_{e|f(D)} = \sigma_\epsilon^2 I_{n_{tot}} + (C_D((\mathbf{x}_{ij}^e,\psi_i),(\mathbf{x}_{i'j'}^e,\psi_{i'})))_{1\le i,i'\le n_e, 1\le j,j'\le n_i}$. Note that $\mathbf{m}_{e|f(D)}$ is a vector of size n_{tot} and $\mathbf{V}_{e|f(D)}$ is $n_{tot}\times n_{tot}$ matrix. This corresponds to what we called the complete mixed meta-model. The likelihood cannot be factorized as a product of individual likelihoods which makes the latent parameters ψ_i all dependent. Moreover, the computation of the likelihood requires the inversion of a $n_{tot}\times n_{tot}$-matrix ($\mathbf{V}_{e|f(D)}$) at each iteration which is highly computationally intensive. Therefore, we propose an intermediate mixed meta-model by replacing the matrix $\mathbf{V}_{e|f(D)}$ by its diagonal. This leads to a likelihood denoted by $\bar{\ell}_D(\theta,\sigma_\epsilon^2;\mathbf{y}^e)$ which is fast to compute. And we neglect totally the additional uncertainty coming from the emulator, we obtain the simple mixed meta-model with the likelihood $\tilde{\ell}_D(\theta,\sigma_\epsilon^2;\mathbf{y}^e)$ where $\mathbf{V}_{e|f(D)}$ is replaced with $\sigma_\epsilon^2 I_{n_{tot}}$. This likelihood is simply $\ell(\theta,\sigma_\epsilon^2;\mathbf{y}^e)$ where m_D was substituted for f.

Algorithm 2: SAEM-MCMC algorithm for the mixed meta-models

At iteration k, given the current values of the estimators $\hat{\mu}^{(k-1)}, \hat{\Omega}^{(k-1)}, \hat{\sigma}_\epsilon^{2(k-1)}$:

Simulation step: For each individual i successively, update $\psi_i^{(k)}$ with m iterations of an MCMC procedure with $\pi(\psi_i|\mathbf{y}_i^e;\widehat{\theta}^{(k-1)},\psi_{-i},\sigma_\epsilon^2)$ as stationary distribution.

Stochastic Approximation step: update the sufficient statistics $s_{k,1}$, $s_{k,2}$ and $s_{k,3}$ following the stochastic approximation scheme ($l=1,2,3$):

$$s_{k,l} = s_{k-1,l} + \gamma_k \left(S_l(\mathbf{y}^e,\psi^{(k)}) - s_{k-1,l} \right)$$

Maximization step: update the population parameters

$$\widehat{\mu}^{(k)} \;=\; \frac{s_{k,1}}{n_e}, \quad \widehat{\Omega}^{(k)} = \frac{s_{k,2}}{n_e} - \frac{s_{k,1}\, s_{k,1}^t}{n_e^2}, \quad \widehat{\sigma}_\epsilon^{2(k)} = \frac{s_{k,3}}{n_{tot}}.$$

An SAEM algorithm (Algorithm 2) is used to estimate the unknown parameters $\theta = (\mu,\Omega)$ and σ_ϵ^2 by maximizing one of the three likelihoods of the mixed meta-models. The simulation step is performed via an MCMC as suggested in [99].

The stationary distribution depends on the considered likelihood. If the likelihood $\ell_D(\theta, \sigma_\epsilon^2; \mathbf{y}^e)$ is considered, when updating the individual parameter ψ_i, the other individual parameters ψ_{-i} shall be taken into account in the likelihood computation. Otherwise, the updata on the ψ_i's can be run in parallel since $\pi(\psi_i|\mathbf{y}_i^e; \widehat{\theta}^{(k-1)}, \psi_{-i}, \sigma_\epsilon^2) = \pi(\psi_i|\mathbf{y}_i^e; \widehat{\theta}^{(k-1)}, \sigma_\epsilon^2)$. The sufficient statistics for the simple mixed meta-model are standard as shown in [148]: $S_1(\mathbf{y}, \psi) = \sum_{i=1}^{n_e} \psi_i$, $S_2(\mathbf{y}^e, \psi) = \sum_{i=1}^{n_e} \psi_i \psi_i^t$ and $S_3(\mathbf{y}, \psi) = \sum_{i,j}(y_{ij}^e - m(\mathbf{x}_{ij}^e, \psi_i))^2$. For the complete and intermediate mixed meta-model, the statistic $S_3(\mathbf{y}, \psi)$ needs to be adapted. The sequence $(\gamma_k)_{k\geq 0}$ is a sequence of positive numbers decreasing to 0 ($0 < \gamma_k \leq 1$).

Under some standard assumptions as the ones done in [99], the SAEM algorithm (Algorithm 2) produces a sequence of parameters converging to the maximum of the corresponding likelihood. Following [56], we need a uniform control decreasing with N (the number of point in the DoNE) on the distances between the approximated likelihoods (ℓ_D, $\bar{\ell}_D$ and $\tilde{\ell}_D$) and the true one ℓ to ensure that the estimates produced by the SAEM algorithm with an approximate likelihood will converge to the maximum likelihood estimates of the true likelihood (Eq 2.18). This is given in the next proposition:

Proposition 2. *Under some standard assumptions close to the ones in Proposition 1, we have*

$$|\ell(\theta, \sigma_\epsilon^2; \mathbf{y}^e) - \hat{\ell}_D(\theta, \sigma_\epsilon^2; \mathbf{y}^e)| \leq \hat{C}st_{\mathbf{y}^e} \frac{n_{tot}}{\sigma_\epsilon^{n_{tot}+2}} G_C(h_D)$$

where $\hat{\ell}_D$ may be any of the approximated likelihood (ℓ_D, $\bar{\ell}_D$ or $\tilde{\ell}_D$), $\hat{C}st_{\mathbf{y}^e}$ is a constant depending on the chosen approximated likelihood and the data, h_D is the covering distance associate with the DoNE and G_C is a function decreasing to 0 with $h_D \to 0$ and is provided in [151] for classical kernel.

The simulation studies have confirmed these theoretical results. In particular, it was noticed that increasing the number of points N in the DoNE makes the estimates closer to the maximum of the true likelihood. Working with the intermediate or the simple mixed meta-model greatly improve the computational time in comparison with the complete likelihood. The major loss was especially for the estimation of the observation noise σ_ϵ^2. This noise is overestimated under the simple mixed meta-model since an additional source of uncertainty (substituted of f by its approximation m_D) is neglected while it may be underestimated if we ignore the covariance of the error structure of $f - m_D$ as done in the intermediate mixed meta-model. Indeed in the latter model, the approximation uncertainty may be confused with the observation noise.

2.2.3 Accounting for Simulator Error

The question at stake in [JP10] and [P3] was to decide whether the simulator can be a good enough representation of the real phenomenon. This question can be transposed to the discrepancy, deciding whether $\delta = 0$ or not. In a statistical decision framework combining Equations (2.5) and (2.4) lead to decide between hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad y_i^e = f(\mathbf{x}_i^e, \theta) + \epsilon_i, \\ \mathcal{H}_1: &\quad y_i^e = f(\mathbf{x}_i^e, \theta) + \delta(\mathbf{x}_i^e) + \epsilon_i. \end{aligned}$$

Under $\mathscr{H}_1$, we assume that $\delta \sim \mathscr{GP}(0,\sigma_\delta^2 C_\delta(\cdot,\cdot))$ and under $\mathscr{H}_0$ and $\mathscr{H}_1$ we still assume that $\epsilon_i \overset{i.i.d.}{\sim} \mathscr{N}(0,\sigma_\epsilon^2)$.

In a Bayesian framework, the decision between two hypotheses may rely on the computation of a Bayes factor [92]. A more recent approach consists in considering a mixture model that encompasses the models defined by the two hypotheses in competition [90]. The former approach was used in [JP10] while the latter was adopted in [P3]. Both rely on a simplification assumption on the simulator, it is assumed that the simulator is linear in θ i.e. $f(\mathbf{x},\theta)=g(\mathbf{x})^T\theta$.

The Bayes factor is the ratio between the two integrated likelihoods of the two models:

$$B_{0,1}(\mathbf{y}^e) := \frac{\pi(\mathbf{y}^e|\mathscr{H}_0)}{\pi(\mathbf{y}^e|\mathscr{H}_1)} \quad \text{where} \quad \pi(\mathbf{y}^e|\mathscr{H}_j)=\int_{\xi_j}\pi(\mathbf{y}^e|\xi_j,\mathscr{H}_j)\pi(\xi_j)d\xi_j .$$

In the expression above, ξ_j denotes all the parameters of the model in $\mathscr{H}_j$, it includes θ and σ_ϵ^2 in both models and also the parameters of the discrepancy for the model in $\mathscr{H}_1$. Compatible priors [39] or objective priors [28] have to be chosen to ensure a *fair* comparison between the two models. However, these priors have to be proper, otherwise the marginal likelihood $\pi(\mathbf{y}^e|\mathscr{H}_0)$ is ill-defined. A solution is to use an intrinsic Bayes Factor [10] where partial Bayes factor are computed by using a part of the data to make proper an improper prior. Then, the intrinsic Bayes factor is obtained as a mean of the partial Bayes factor for all possible partitions between data used for making the prior proper and data on which the marginal likelihood is computed. In the model comparison at hand, under simple assumptions on the prior distribution, the following proposition gives an identity between the intrinsic Bayes factor and the standard Bayes factor. Under hypothesis $\mathscr{H}_1$, we consider the parameters $(\sigma_\delta^2, k=\sigma_\epsilon^2/\sigma_\delta^2)$ instead of $(\sigma_\delta^2,\sigma_\epsilon^2)$ in addition to θ and σ_ϵ^2.

Proposition 3. *If $\pi(\xi_0)=1/\sigma_\epsilon^2$, $\pi(\xi_1)=\pi(\theta,\sigma_\delta^2,\psi,k)=\pi(\psi|k)\pi(k)/\sigma_\delta^2$ with $\pi(\psi,k)$ proper and $m=d+1$ then the intrinsic Bayes factor is:*

$$B_{0,1}^A(\mathbf{y}^e)=\frac{B_{0,1}(\mathbf{y}^e)}{C_{n,n_0}}\sum_{|\mathscr{A}|=n_0} B_{0,1}(\mathbf{y}^e(\mathscr{A}))^{-1}=B_{0,1}(\mathbf{y}^e)$$

where C_{n,n_0} is the number of choices of n_0 items among n items and $B_{0,1}(\mathbf{y}^e(\mathscr{A}))$ is the partial Bayes factor where the subset $\mathscr{A}$ is used to make the prior proper.

Another solution is to consider a mixture of the two models under the two hypotheses in competition. We denote respectively by $\ell_{\mathscr{H}_0}$ and $\ell_{\mathscr{H}_1}$ the likelihoods of the models under the two hypotheses and by $\mathscr{M}_\alpha$ the mixture model: for $i=1,\ldots,n_e$,

$$\mathscr{M}_\alpha : y_i^e \overset{ind.}{\sim} \alpha\left(\ell_{\mathscr{H}_0}(\theta,\sigma_\epsilon^2;y_i^e)\right)+(1-\alpha)\left(\ell_{\mathscr{H}_1}(\theta,\sigma_\epsilon^2,\delta;y_i^e)\right).$$

This mixture model has to be written conditionally on the discrepancy δ in order to make the y_i^e independent. We provided a theorem in [P3] which guarantees that the posterior distribution of the parameters is proper for a Jeffreys prior on $(\theta,\sigma_\epsilon^2)$ and a proper prior on the other parameters. The inference was conducted through a Metropolis-within-Gibbs algorithm where all the parameters are updated successively. This algorithm also simulates conditionally on the parameters, the discrepancy at the

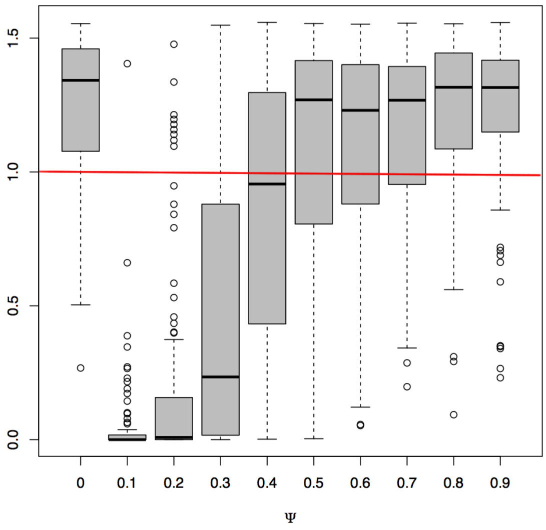


Figure 2.10 – *Bayes factor computations over* 100 *simulated datasets of size* $n_e = 30$. *The x-value gives the range parameters of the exponential kernel used to generate the discrepancy. The simulator is a* $f(\mathbf{x}) = (1, x, x^2)(\theta_0, \theta_1, \theta_2)^t$.

locations $\mathbf{x}^e$ and binary variables for each $i \in \{1, \ldots, n_e\}$ indicating whether the data y_i^e was generated under $\mathscr{H}_0$ or $\mathscr{H}_1$.

In [JP10], a simulation study was conducted to investigate the ability of the Bayes factor to determine which model is the most consistent with the data. The artificial data were generated by using a Gaussian process to generate the discrepancy. Different values for the correlation range of the correlation of the GP were tested. Some results are reproduced in Figure 2.10. A similar study was conducted in [P3]. Some results are reproduced in Figure 2.11. Both methods are able to detect the absence of discrepancy (results not reported here for the mixture model). In Figure 2.11, if the correlation range parameter is too small the discrepancy is not distinguishable from a white noise. This is also expected with the intrinsic Bayes factor. What is more surprising is that the two methods favor the zero discrepancy model when the correlation range parameter is large. This is a result of the confounding effect which makes the discrepancy to have a smoothness similar to the simulator. A value of θ different from the true one θ^*, may compensate for the discrepancy. More precisely, another value $\tilde{\theta}$ makes the simulator close to the simulator with the true θ^* plus discrepancy: $f(\mathbf{x}, \tilde{\theta}) \approx f(\mathbf{x}, \theta^*) + \delta(\mathbf{x})$ for all $\mathbf{x}$. This is illustrated with both methods.

2.2.4 Post-processing probabilistic meteorological and hydrological forecasts

The papers [JP3] and [JP9] deal with post-processing of meteorological forecasts, respectively temperature - precipitation forecasts and water flow forecasts. A probabilistic forecast consists in a probability distribution of a given quantity of interest in the future, for instance the amount of precipitation or the mean temperature the day to come. This distribution has to be well calibrated and sharp [72]. In this context, calibration has a different meaning than in the usual framework of UQ. Indeed, calibration means that the prediction intervals at a given level of confidence derived from the prob-

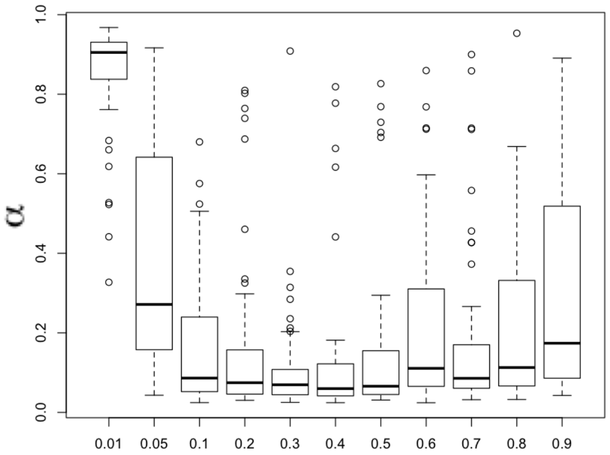


Figure 2.11 – *Posterior mean estimates of the mixture parameter α over simulated* 100 *datasets of size $n_e = 50$. The x-value gives the range parameters of the exponential kernel used to generate the discrepancy. The simulator is a $f(\mathbf{x}) = (1, x, x^2)(\theta_0, \theta_1, \theta_2)^t$.*

abilistic forecast should effectively meet this level of confidence. Sharpness means that the widespread of the probabilistic distribution should be as narrow as possible.

In meteorology, the forecasts rely on ensemble forecasting. The members of an ensemble are the outputs of complex meteorological simulator for which the inputs giving the initial conditions have been slightly perturbed. Meteorologists would like to consider this ensemble as a sample of the probabilistic forecast. However, this distribution is often shown to be over confident which supports the need for post-processing to achieve a better calibration. In [JP3], we proposed an exchangeable model to post-process the temperature and precipitation forecasts from several ensembles.

When the issue at stake is the production of hydroelectricity, the probabilistic forecast of interest is the water flow of a river. This forecast is obtained by inputting temperature and precipitation forecast into a conceptual simulator named a rainfall-runoff model which outputs the water flow of a river. Even if true temperature and amount of precipitation are feeding in the simulator, the simulator may still suffer from some discrepancy. This discrepancy strongly depends on the hydrological regimes: rapid flood variations induce large errors of anticipation but a series of dry events will translate into a much more smoother sequence of river levels due to the easily predictable behavior of the soil reservoir emptying. That is why we proposed in [JP9], a two-regime statistical model which embeds the runoff-rainfall simulator. The two regimes correspond to different discrepancy structures. The river regime is modeled as a latent variable, the distribution of which is based on additional outputs of the rainfall-runoff simulator.

These two contributions in the field of meteorological forecast may lead to more general contributions. In particular, the combination of several simulators in a probabilistic forecasting goal is interesting for other fields than meteorology. Modeling discrepancy with several working regimes depending either on the state of the nature or on some particular working conditions of the simulator could be also sensible in many applications.

2.2.4.1 Post-processing meteorological forecasts

The model we proposed in [JP3] relies on Gaussian distribution assumptions. Up to a Box-Cox transformation, this assumption is reasonable for temperatures. To deal with precipitations, we introduce pseudo-precipitations which are assumed to be normally distributed. They correspond to real precipitations if they are positive and are latent otherwise.

Let B be the number of forecast sources (e.g. the ensembles from several meteorological centers) and K_b the number of members within ensemble b. The members from an ensemble b for forecasting time t are denoted by $(y_{b,k,t})_{k=1,\dots,K_b}$. They are obtained by perturbing the initial condition of a meteorological simulator at time $t-h$, h being the time horizon of the forecast. The real meteorological quantity to be forecasted is y_t^e. We aim to link the meteorological quantity of interest to the forecast sources. We make the assumption of exchangeability within an ensemble which leads to the existence of a latent variable [53] accounting for the shared information between the members. Moreover, we assume that this latent variable is common to all the ensembles and that another common latent variable accounts for the dispersion of the member. More precisely, the model we propose is for a given time horizon and a given location: for any b,k,t,

$$\begin{cases} \left(y_{b,k,t}|U_t\right) & = \alpha_b + \beta_b U_t + \gamma_b \varepsilon_{b,k,t} \\ (y_t^e|U_t) & = a_0 + U_t + \varepsilon_{0,1,t} \\ \left(\varepsilon_{b,k,t}|V_t^{-2}\right) & \overset{ind}{\sim} \mathcal{N}\left(0, V_t^2\right) \\ \left(U_t|V_t^{-2}\right) & \overset{ind}{\sim} \mathcal{N}\left(0, \lambda V_t^2\right) \\ V_t^{-2} & \overset{iid}{\sim} \Gamma(\alpha_\Gamma, \beta_\Gamma) \end{cases}, \tag{2.20}$$

where U_t and V_t^2 are the corresponding latent variables (forming the bedrock of the exchangeability property) upon which the ensemble members of a given ensemble b are conditionally independent. These latent variables U_t and V_t^2 are assumed to be independent across time. The parameters α_Γ, β_Γ, λ and $\{\alpha_b, \beta_b, \gamma_b\}_{b\in\{0,\dots,B\}}$ are parameters to be estimated. These parameters are then specific to the considered time horizon and location. Identifiability constraints impose $b_0 = c_0 = 1$. The parameters are to be interpreted as:

- The difference $\alpha_b - \alpha_0$, $b > 0$ gives the additive bias for the forecasting ensemble v, to be compared to 0.

- The ratio $\frac{\beta_b}{\beta_0}$, $b > 0$ is the multiplicative bias of the forecasting ensemble b. Since $\beta_0 = 1$ for identifiability, the value β_b is directly to be compared to 1. Additive and multiplicative biases may partly compensate one another.

- For parameter γ, the ratio $\frac{\gamma_b}{\gamma_0} = \gamma_b$, $b > 0$ (parameter γ_0 being fixed to 1) will be understood as a dispersion bias for the predictors. A ratio greater than 1 can be interpreted as an over-dispersion of the predicting ensemble b .

- The ratio $\frac{\beta_\Gamma}{\alpha_\Gamma - 1}$ corresponds to the expected value of V_t^2 which rules how far the quantity to forecast y_t^e can occur from the latent variable U_t. It is therefore expected that this ratio will increase with the time horizon of the forecast, because ensembles generally become less and less informative when the forecasting horizon grows.

The *adhoc* dependence between U_t and V_t^2 as specified by Eq 2.20 greatly facilitates inference and forecasting (through the property of Gamma-Normal conjugacy) and therefore leads to fast algorithms, which is useful in an operational context, where inference can be conducted within a moving window. Inference relies on an EM algorithm where the Expectation step has a close form for the temperature and requires stochastic simulation in the Expectation step for the pseudo precipitation.

This method was tested on real watersheds with data provided by Hydro-Québec. Most results showed that this model which allows us to post-process multi-ensemble data, led to better probabilistic forecast according to the CRPS score [72]. As an illustration, Figure 2.12 shows the post-processed forecast of the maximal daily temperature from three forecasting ensembles. On this particular example, the post-processed forecast manages to make a trade-off between the different ensembles and produces a probabilistic forecast that covers with high probability the true maximal temperature. We can also observe that, as expected, the dispersion increases with the time horizon.

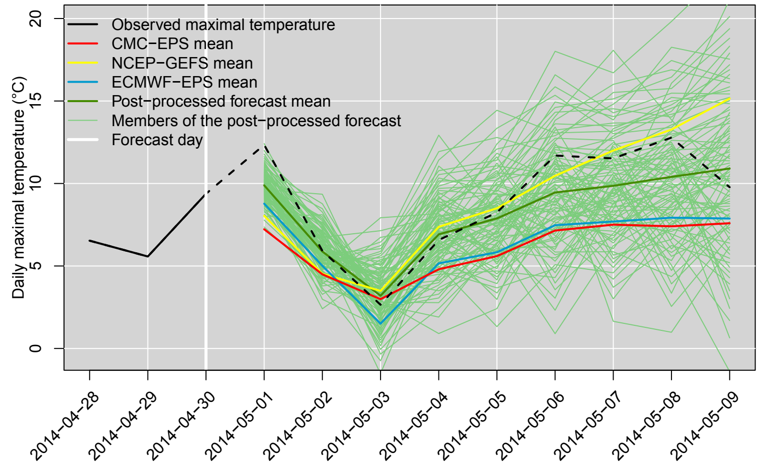


Figure 2.12 – *Example of a forecast issued for the daily maximum temperatures of the Manic 2 catchment area with the proposed post-processing method taking NCEP-GEFS, CMC-EPS and ECMWF-EPS as inputs. Predictive scenarios derived from meteorological forecasts by forecast time horizon are presented. The maximum daily temperature to be forecast (observed afterward) is indicated by the black dotted line.*

2.2.4.2 A two-regime model for rainfall-runoff simulator (RRS)

From the data we have on watersheds in Québec, provided by Hydro-Québec, and in France, provided by EDF, we compared the outputs of the RRS (where the actual temperatures and precipitations are inputted) with the measured water flow. We frequently observed two situations that suggest a model with two regimes: (i) a situation where the RRS ingests new rainfalls, the link between the RRS output and the waterflow is subjected to a high level of uncertainty, (ii) a situation where there is no recent rain event to be inputted in the RRS, the time increments of the observed water flow and of the outputs of the RRS are very close to each other, the actual error is propagated identically from one time step to the next step. Consequently, the prediction uncertainty is much smaller. The RRS is a conceptual model which mimics the river behavior by a system of interconnected reservoirs. Besides the water flow of the river, internal state variables (such as snowmelt intensity, fraction of the superficial flow in the whole flow, etc.) are also available and may bring valuable information about the regimes of the river.

In this subsection, we denote respectively by y^e_t and y_t the logarithms of actual water flow observation and of the output of the RRS at day t. We used a logarithm transformation to make Gaussian distribution assumption acceptable for the data. We would like our model to properly identify different regimes such as the ones presented above and to predict which one is adequate for any prediction situation. We propose to make the conditional distribution of $\left(y^e_t | \mathbf{y}_t, \mathbf{y}^e_{t-1}\right)$ depend on the sign of a latent variable denoted by U_t indicating the regime of the river and thus the nature of the relationship between y^e_t, $\mathbf{y}_t$ and $\mathbf{y}^e_{t-1}$. We assume that such a latent variable follows a Gaussian distribution, the mean of which is a function of the RRS state variables at the considered time, $\mathbf{V}_t$. Thus, this model exhibits two regimes that we name regime 0 and regime 1. The regime is given by the binary random variable $S_t = \mathbb{I}_{\{U_t \leq 0\}}$ (where $\mathbb{I}_{\{A=a\}}$ is the indicator function of the event $\{A = a\}$) which says whether the RRS behaves according to regime 0 or regime 1 at time t. Let $\mathbf{V}_t$ denote the vector of state variables of the RRS at time t and we assume that:

$$U_t \underset{i.i.d.}{\sim} \mathcal{N}(\mathbf{B}^T \mathbf{V}_t, 1), \tag{2.21}$$

where the vector $\mathbf{B}$ contains unknown parameters to be estimated. We emphasize that we only assume that the U_ts (hence the S_ts) are independent conditionally to the state variables $\mathbf{V}_t$. Hence, the regimes at time t and $t+1$ are still dependent since $\mathbf{V}_t$ and $\mathbf{V}_{t+1}$ are dependent.

We moreover assume that, conditionally on $(S_t = k)$, $(\mathbf{y}^e | \mathbf{y})$ behaves according to an autoregressive model with external inputs (ARX) model given by the following equation:

$$y^e_t = a_k + \mathbf{b}_k^T \mathbf{y}_t^{t-r} + \mathbf{c}_k^T \mathbf{y}_{t-1}^{e,t-s} + \sigma_k \varepsilon_t \quad \varepsilon_t \underset{i.i.d.}{\sim} \mathcal{N}(0,1) \tag{2.22}$$

where $\mathbf{y}_{t-1}^{e,t-s} = (y^e_{t-1}, y^e_{t-2}, \ldots, y^e_{t-s})^T$, $\mathbf{y}_t^{t-r} = (y_t, y_{t-1}, \ldots, y_{t-r})^T$ and $\theta = (\theta_k)_{k\in\{0,1\}} = \left(a_k, \mathbf{b}_k^T, \mathbf{c}_k^T, \sigma_k\right)_{k\in\{0,1\}}$ stands for the set of unknown parameters. The parameters $\{a_k, \sigma_k\}_{k\in\{0,1\}}$ and vectors of parameters $\{\mathbf{b}_k, \mathbf{c}_k\}_{k\in\{0,1\}}$ are to be estimated (in addition to the vector $\mathbf{B}$). Whatever the regime k, $\mathbf{b}_k$ is of size $r+1$ and $\mathbf{c}_k$ is of size s. Thus, the statistical modeling of the link between the outputs of the RRS, $y_t, y_{t-1}, \ldots$, and the actual water flow, $y^e_t, y^e_{t-1}, \ldots$, has two regimes ($S_t = 0$ or $S_t = 1$) which depend on the latent variable U_t governed by the known state variables $\mathbf{V}_t$.

The inference is conducted through an EM algorithm with exact Expectation steps. A model selection relying on a Bayesian Information Criterion (BIC) [153] is used for selecting the indices s and r in Equation (2.22) and for selecting the state variables to keep in Equation (2.21).

This model was inferred on 11 French watersheds and 4 Québec watersheds. For most watersheds, from the interpretation of the parameter estimates, we identify:

- a regime (by convention $S = 0$) in which approximately only the time increments of $\mathbf{y}$ and $\mathbf{y}^e$ are considered (i.e. $b_{10} \simeq -b_{20}$ and $c_{10} \simeq 1$ in Equation (2.22)),

- a regime (by convention $S = 1$) in which the predicted water flow depends more on the RRS outputs ($c_{11} < c_{10}$ and $b_{11} > b_{10}$), approximately only the errors

$(\mathbf{y}^e - \mathbf{y})$ matter ($b_{11} \simeq 1$ and $c_{11} \simeq -b_{21}$) and with a higher uncertainty level ($\sigma_1 > \sigma_0$).

We also identified for the different watersheds which state variables are responsible for being in a state or the other. As an illustration, we provide an illustration in Figure 2.13 of the flow prediction on the Dordogne watershed at Bort. We compare the probabilistic forecast we obtain with the two-regime model with a one-regime model (same model as in Equation (2.22) without the k index, i.e. without the dependence on the state variables) and with an operational prediction method used by EDF. We obtain a water flow prediction closer to the actual one. We notice that the probability for being in regime 1 rather than 0 increased when there was a variation in the water flow and then vanished when the water flow goes back to steady state.

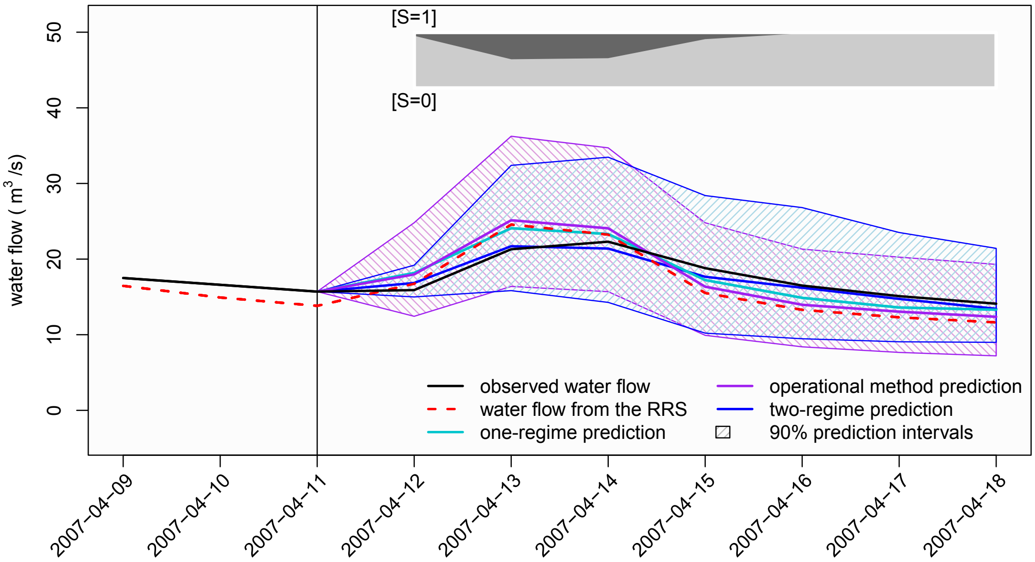


Figure 2.13 – *Water flow prediction on the Dordogne watershed at Bort. The proportion of the dark gray filling of the upper right box corresponds to the predicted probability of being in the high uncertainty (flood) regime ($S = 1$). The vertical line corresponds to the date until which one has access to the water flow observations (i.e. the forecasting day). The prediction targets the water flow observations after that day.*

2.3 Perspectives

My perspectives are mainly devoted to pursue the investigation regarding the discrepancy function and to extent some classical techniques of UQ for stochastic simulators. In the different perspectives, the question of the design of field or numerical experiment is at stake but with respect to different objectives. Some perspectives are already ongoing works while others are longer-term projects.

2.3.1 Discrepancy

When confronting field data to a simulation, a systematic error between the real phenomenon and the simulator is usually taken into account as in Equation 2.5. To model an unknown function, the Gaussian process assumption makes sense. Under this assumption, running a variable selection of input variables in the vector $\mathbf{x}^e$ will help to detect which input variables significantly impact the simulator. In [JP10] and [P3], under a simplifying assumption on the simulator, we proposed Bayesian testing procedure to decide whether the discrepancy is required or not. The extension of the testing procedure when removing this assumption is a natural perspective. Moreover, other models than a Gaussian process for the discrepancy could be tested. When the simulator is the solution of differential equations (often available as a numerical solver of the differential equations), a discrepancy can be embedded directly within the equations which may make extrapolation possible.

Screening the model discrepancy. Analyzing the discrepancy should help to understand to what extent the simulator is reliable. In particular, we focus on determining whether some variables are active or inert in the discrepancy function, which is of major interest since it indicates which input variables are correctly taken into account in the simulator. Therefore, this could give some leads to improve the simulator and help to determine whether extrapolation is safe or not with respect to a specific input. The major difficulty in selecting the active variables in the discrepancy function is that the discrepancy function is not directly observed and confounding effects with the calibration of the simulator can occur.

We consider this model for the discrepancy: $\delta \sim \mathcal{GP}(0, \sigma_\delta^2 C_\delta(\cdot,\cdot))$ where C_δ is restricted to the class of power exponential kernels:

$$C_\delta(\mathbf{x}, \mathbf{x}'; \alpha, \psi) = \prod_{i=1}^{p} \exp(-|x_i - x_i'|^\alpha / \psi_i)$$

which encompasses the exponential and the Gaussian kernels. The parameter α is usually chosen by the user and not estimated from data. The parameters $\psi_i \in (0; +\infty)$ are named the range parameters and correspond respectively to the input variables x_i. Note that if $\psi_i \to +\infty$, the corresponding input variable does not have any impact on the discrepancy and can be deemed as inert. Another parametrization is used in [111] for the range parameters which plunges the range parameters into the unit interval $[0, 1]$. It is given by $\rho = \exp(-2/\psi)$ which leads to the correlation kernel:

$$C_\delta(\mathbf{x}, \mathbf{x}'; \alpha, \rho) = \prod_{i=1}^{p} \rho_i^{2^\alpha |x_i - x_i'|^\alpha}. \tag{2.23}$$

Then an inert input variable corresponds to $\rho_i = 1$. In [111], a spike and slab prior [67] is set on the range parameters ρ_i:

$$\pi(\rho) = \prod_{i=1}^{p} \Big[\tau I_{[0,1]}(\rho_i) + (1-\tau) d_1(\rho_i) \Big] \tag{2.24}$$

where $\tau \in [0, 1]$ is the prior probability that an input variable is active (common to all i's) and $d_1(\cdot)$ is a Dirac mass at 1. The prior in (2.24) can be obtained by stating that

$$\begin{aligned} \pi(\rho \mid \gamma) &= \prod_{i=1}^{p} \Big[\gamma_i \, I_{[0,1]}(\rho_i) + (1-\gamma_i) \, d_1(\rho_i) \Big] \\ \pi(\gamma) &= \prod_{i=1}^{p} \tau^{\gamma_i} (1-\tau)^{1-\gamma_i} \end{aligned} \tag{2.25}$$

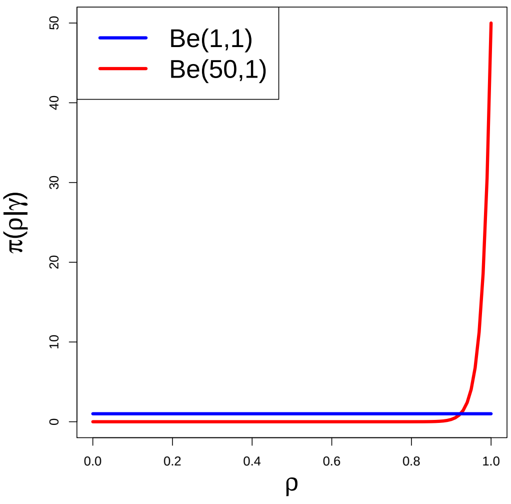


Figure 2.14 – *Beta distribution in the mixture of the smooth spike and slab prior for* $\alpha = 50$.

and then marginalizing over γ. In [150], the authors use (2.25) and devise MCMC schemes to sample from the joint posterior $\rho, \gamma \mid \mathbf{y}^e$ for fixed τ, which makes performing the screening selection exercise straightforward — all the relevant information is in the posterior distribution of γ. Note that the contribution of [111] is concerned with the emulation by a GP of a simulator and the contribution of [150] deals with GP regression. None of them tackles the variable selection in the discrepancy function when the calibration of the simulator may have to be conducted jointly.

Our proposition is to take a smooth spike and slab prior as:

$$\pi(\rho \mid \gamma) = \prod_{i=1}^{p} [\gamma_i \operatorname{Be}(\rho_i \mid 1,1) + (1-\gamma_i) \operatorname{Be}(\rho_i \mid \alpha_i, 1)] \tag{2.26}$$

where $\operatorname{Be}(\cdot \mid \alpha, \beta)$ is the Beta density with positive shape parameters α and β. In Equation (2.26), α_i is a large value, typically larger than 50. Thus, for all input dimensions, the prior distribution is a mixture of a uniform distribution ($\operatorname{Be}(\cdot \mid 1,1)$) which may correspond to an active input and of a distribution highly concentrated around 1 (instead of being a Dirac mass at 1 as in the discrete spike and slab) corresponding to an inert input. The two parts of the prior are depicted in Figure 2.14. We have the intuition that as $\alpha_i \to +\infty$, the resulting inference approaches the one obtained via the discrete spike and slab prior, so that this construction can be viewed as a relaxation technique to facilitate the sampling. By doing so, the prior distribution is absolutely continuous with respect to Lebesgue measure which is not the case with the discrete spike and slab.

The choice of the vector $\gamma \in \{0,1\}^p$ leads to 2^p models in competition differing only on the prior distribution for ρ. To determine which are the active or inert variables, we need to compute the posterior distribution for each model. These posterior distributions are obtained up to a constant by computing the Bayes factors of each competing model to the full model, $\gamma = \mathbf{1}$, which we denote by

$$B_\gamma = \frac{m(\mathbf{y}^e \mid \gamma)}{m(\mathbf{y}^e \mid \mathbf{1})}, \tag{2.27}$$

where $m(\mathbf{y}^e \mid \gamma) = \mathbb{E}(\ell(\rho, \eta; \mathbf{y}^e)\ \pi(\rho, \eta \mid \gamma))$ with $\ell(\rho, \eta \mid \mathbf{y}^e)$ the likelihood of the model defined by Equations (2.4) and (2.5), and η the other parameters than the range parameters ρ (it includes the calibration parameter θ and the variance parameters σ_ϵ^2 and σ_δ^2).

Since the parameter spaces of the models in competition are the same, we can estimate B_γ by using a version of importance sampling [36] as an expectation under $\mathcal{M}_1$

$$B_\gamma = \mathbb{E}_1\left[\frac{\ell(\rho,\eta;\mathbf{y}^e)\,\pi(\rho,\eta\mid\gamma)}{\ell(\rho,\eta;\mathbf{y}^e,)\,\pi(\rho,\eta\mid\mathbf{1})}\right] = \mathbb{E}_1\left[\frac{\pi(\rho\mid\gamma)}{\pi(\rho\mid\mathbf{1})}\right], \tag{2.28}$$

As the only difference between two models is the prior distribution on ρ the expectation simplifies. All the Bayes factor can be estimated from a unique MCMC sample from the model $\mathcal{M}_1$: if $\{(\rho^{(k)},\eta^{(k)}),\ k=1,\dots,M\}$, is a sample from the posterior distribution of the unknowns for $\gamma=\mathbf{1}$, then

$$B_\gamma \approx \frac{1}{M}\sum_{k=1}^{M}\pi(\rho^{(k)}\mid\gamma).$$

Therefore, the posterior probability for any input variable to be active in the discrepancy is computed as:

$$\mathbb{P}(x_k \text{ active in } \delta\mid\mathbf{y}^e) = \sum_{\gamma:\,\gamma_k=1}\mathbb{P}(\mathcal{M}_\gamma\mid\mathbf{y}^e). \tag{2.29}$$

We applied the developed method to the PV simulator detailed in [JP2] and recalled in Subsection 2.2.2.1. We considered 5 input variables (the four variables which are taken into account in the simulator and another measure of the temperature which is available in the field experiments). As Figure 2.15 illustrates, the posterior probabilities for most of the variables to be active in the discrepancy are high except for the two temperatures. Nevertheless, at least one of them should be incorporated within the discrepancy since the probability for either a temperature or the other is high. This work is in collaboration with Rui Paulo and Anabel Forte and should be submitted soon for publication [IP1]. It assumes that the simulator is cheap or that we are already working with a surrogate of the simulator. But it should be extended to the case where an emulator has to be used instead of the simulator.

Testing discrepancy. The natural extension of [JP10] and [P3] is to remove the linear assumption of the simulator with respect to the calibration parameter θ. If the simulator is not linear, both methods can still be used with a linear surrogate of the simulator. However, the validation which is done by deciding whether the discrepancy should be or not incorporated concerns the linear surrogate and not directly the complex simulator. Removing the linear assumption will lead to additional computational burden, since it will not be possible to integrate out the calibration parameter θ. These methods could also be used to decide between different types of discrepancy. It could be different Gaussian processes with different correlation kernels, a different parametric expression or different constraints on the discrepancy [25].

The mixture model in [P3] could also allow us to detect different subregions in the input domain with different kinds of discrepancy. This could be an alternative to the Bayesian treed calibration [97] which gives a partition of the input space, based on a binary tree partitioning, into subregions where different calibration models (as in Equation (2.6)) are fitted. Another alternative for considering different forms of discrepancy depending on the input space was the statistical model proposed in [JP9]

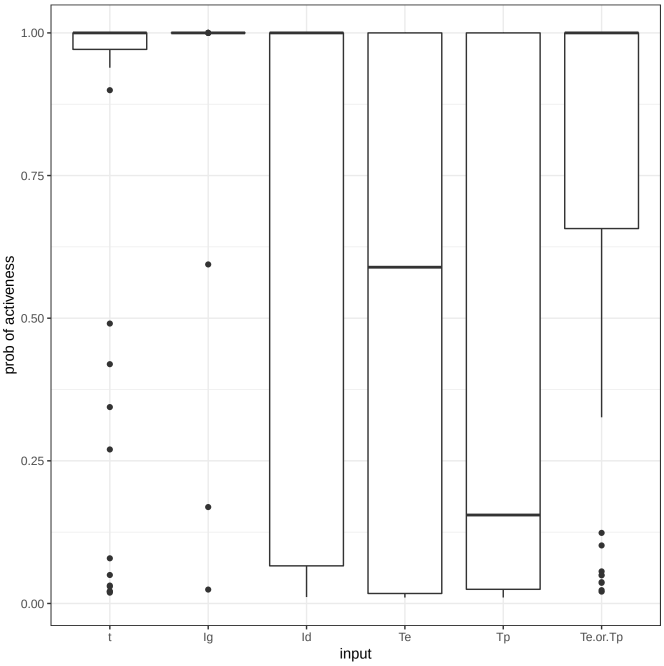


Figure 2.15 – *Posterior probabilities distribution of the activeness of variables in the discrepancy function. The variables are t time, I_g global solar irradiation, I_d diffuse solar irradiation, T_e and T_p temperature at the ground level and temperature on the panel. The last boxplot correspond to the posterior probability distribution that at least one of the two temperatures is active in the discrepancy.*

and recalled in Subsection 2.2.4.2 in the particular context of dynamic observations and simulations. These questions in spite of their difficulty are major since it makes sense to assume that different regimes do exist in the physical process and the simulator may not have the same precision for any of these regimes. To some extent, this question relates to extrapolation of the simulator.

Embedded discrepancy. In the Kennedy and O'Hagan calibration framework [93], the discrepancy is additive in order to model the mismatch between the simulator and field data. An alternative is to embed the discrepancy within the simulator. This makes sense when the simulator is derived from ordinary or partial differential equations (ODE or PDE). It consists in doing a stochastic relaxation of the differential equation(s) which turns the ODE into Stochastic Differential Equations (SDE). We explore this idea for a mass-spring-damper system as in [132]. The position of the mass at time t denoted by $f^R(t)$ is considered as the physical phenomenon of interest and is assumed to follow this ODE:

$$m\ddot{f}^R + c(T)\dot{f}^R + kf^R = 0 \quad \text{with} \quad \dot{T} = c(T)\dot{f}^{R^2} - \frac{1}{\tau}(T - T_0)$$
$$c(T) = \exp\left(\frac{T_0}{T} - 1\right), \quad f^R(0) = 4, \quad \dot{f}^R(0) = 0, \quad T_0 = 20, \quad k = 3, \quad \tau = 1.$$

The mass m and the time t are the input variables. The "field data" $\mathbf{y}^e$ are generated under this ODE at times $0, 1, 2, \ldots 8$ for a mass $m = 1$, then $\mathbf{x}^e = (m = 1, t = 0, \ldots, 8)$. A Gaussian white noise is added, its variance is 10^{-4}. The simulator at hand f is assumed to come from a simplified version of this ODE where c the damper effect is constant:

$$m\ddot{f} + c\dot{f} + kf = 0, \quad f(0) = 4 \quad \dot{f}(0) = 0. \tag{2.30}$$
$$\tag{2.31}$$

Then, the parameters to calibrate are $\theta = (k, c)$ the spring constant and a constant damper effect from the data $(\mathbf{x}^e, \mathbf{y}^e)$. The simplification of the simulator introduces a mismatch between the field data and the outputs of the simulator, especially in an extrapolative setting where the goal is to predict the position (or the speed) of the mass for a larger mass than the one in the field data. If the goal is to predict the position of the mass at the same times $t = 0, \ldots, 8$ and for $m = 2$, we compare four ways of calibrating the simulator. The first calibration consists in considering Equation (2.6) with no discrepancy ($\delta = 0$) and the true noise variance known. The second calibration consists in relaxing, in the first calibration, the known variance assumption. It actually corresponds to model δ as a Gaussian white noise. A third calibration is the method proposed in [132] where a distribution is set on c and the parameters of this distribution are calibrated. A lognormal distribution is set i.e. $c \sim \log\mathcal{N}(\mu_c, \sigma_c^2)$ and the calibration parameters are (μ_c, σ_c^2, k). The fourth and last calibration relies on a transformation of Equation (2.30) into an SDE. It results from a stochastic perturbation on the parameter c for which a practitioner suspects the constant assumption. The stochastic Perturbation is $c = c_0 + \sigma_c \xi_t$ where $\xi_t \sim \mathcal{N}(0, 1)$. The simulator f is then a solution of $\dot{f} = \frac{df}{dt}$ and we have:

$$\begin{aligned} d\dot{f} &= \frac{c}{m}\dot{f}dt + \frac{k}{m}xdt = \frac{c_0 + \sigma_c \xi_t}{m}\dot{f}dt + \frac{k}{m}xdt \\ d\dot{f} &= \frac{c_0}{m}\dot{f}dt + \frac{k}{m}xdt + \frac{\sigma_c}{m}\dot{f}dW_t \end{aligned}$$

where we denote $dW_t = \xi_t dt$ to make the connection with SDE. The value of σ_c is fixed to 0.5.

In Figure 2.16, we plot the prediction performances for mass $m = 2$ of the four calibrations. The predictions are compared with data generated under the true system. The first calibration leads to predictions with really low uncertainties which do not cover the true data. This is really dangerous in practice since it gives the illusion of precision with false predictions. The simple relaxation on the noise gives rather good results. The relaxation proposed by [132] has the desirable feature to increase the credible bands. However the posterior prediction seems rather odd in comparison with the physical behavior of the system. The SDE relaxation (Calibration 4) is promising. All the points have a good validation metric γ and the uncertainties are lower than with the previous relaxation. Moreover, the system behavior seems to be more physically grounded. This was obtained for a good value of σ_c. It is still an open question to tune this parameter properly. A solution could be to use some cross validation techniques.

In the example, the stochastic relaxation is rather natural since the ODE is linear with respect with the parameter c which is relaxed. Another embedding of the discrepancy considers the Voight's failure model [169], where the stochastic relaxation could be done by adding another layer of differentiation. This is a joint work [IP2] with E. Bruce Pitman. From these two examples, we aim to derive general guidance for embedding discrepancy in the simulator. This will enable extrapolation in two ways: outside the domain of field observation and on different outputs (quantity of interest) of the

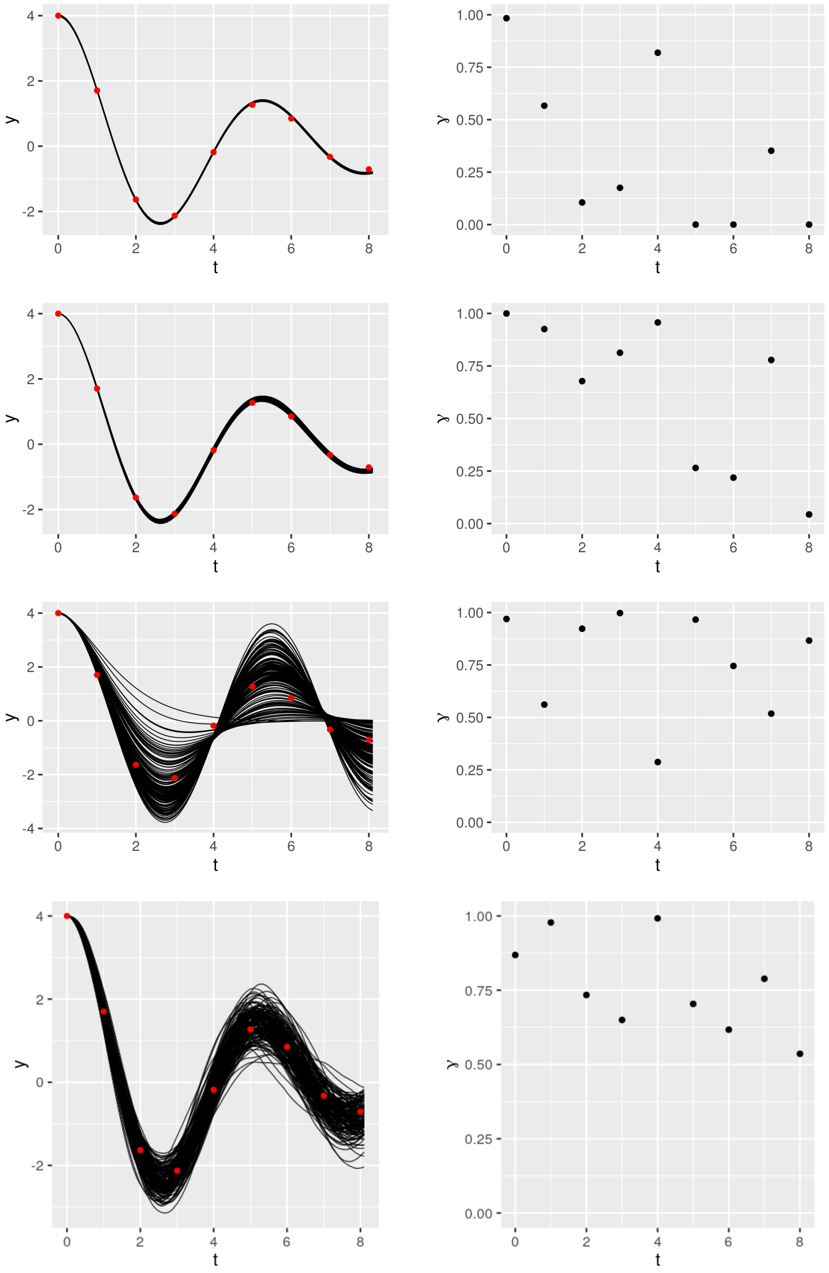


Figure 2.16 – *Top to down, results for calibrations 1 to 4 as described in the text. On the left-hand-side, predictions for the position of the mass ($m = 2$). The predictions correspond to the simulator run with parameters sampled in their posterior distribution. Red points correspond to data generated under the true physical system. On the right-hand-side, computations of the validation metric γ (see Equation (2.7)) for the nine data points generated under the true system.*

simulator than the one available in field observation. The connections with similar ideas in [11] and in [126] need to be done.

2.3.2 Stochastic Simulator

Stochastic simulators are becoming more and more popular in the last decade. The stochasticity within the simulator is either a consequence of stochastic approximations in the simulator computations such as Monte Carlo methods or accounts for the natural stochasticity of the real world process. Most works in UQ make the assumption that the simulator is deterministic and the stochasticity is a result of uncertainties on the input for instance. They need to be extended to deal with stochastic simulators. In the recent years, some works contributed to this area. It seems then important to write a state-of-the-art review of the literature which allows us to identify new research opportunity in this field [P1]. To illustrate the review paper, I dealt with a simulator modeling the concentration of oxygen in the Ocean. The simulator is stochastic since a non-linear PDE is approximated by a Feynman–Kac representation [84]. I considered emulation and calibration of this simulator relying on homoskedastic and heteroskedastic Gaussian Processes [15, 16]. In heteroskedastic Gaussian Process, the mean and the variance of the simulator are both approximated by Gaussian Processes. The R codes are available at: `https://github.com/Demiperimetre/Ocean`.

2.3.3 New Developments for Case Studies

In my contributions, I worked with different simulators. In particular two of them, NitroScape and WALTer, both described in Section 2.2.1, are highly complex and lead to challenging questions. They have in common to be time consuming to run and to have high dimensional inputs and outputs.

Perspectives for NitroScape. NitroScape results from a coupling of 4 simulators iteratively run at the day level. Although some works do exist on building emulators for coupled simulators [100, 119, 118], dealing with a daily-iteration of the simulators may jeopardize the efficiency of these approaches. Then, adapting these emulations is an issue for further work on this simulator including calibration.

Perspectives for WALTer. The outputs of WALTer are comprehensive tillering and Grean Area Index (GAI) dynamics which have to be compared with scarce field data when it comes to calibration. Selecting what features of the outputted dynamics may be collected in field experiments and are sufficient for calibration is an ongoing perspective. Moreover, WALTer is an essential tool for investigating the behavior of a mixture of two wheat varieties sown in the same plot. In this study, the two wheat varieties are assumed to have different phenotype such as their respective height. The quantity of interest is the overyielding that is to say the comparison between the mean number of ears for a mixture of varieties versus the average of mean numbers for the two varieties grown separately. First, the goal will be to realize a sensitivity analysis on this overyielding with respect to the phenotype in order to detect which of them have particular behavior when they are mixed. Since the overyielding is computed from three outputs of WALTer, designing an appropriate DoNE is an issue. Second, we will focus on the best composition of the mixture to optimize the yield.

2.3.4 Designs of Experiments

Since data are scarce because of experimental cost for field data and computation time for simulator runs, choosing them properly is important. This choice should be related to the intented goal. Usually, a first static design is needed and the next experiments or runs may be made sequentially conditionally to the obtained data. Hereafter, I expose particular contexts where some contributions would be relevant.

DoNE for calibration of stochastic simulator. A natural extension of the sequential design for calibration [JP6] is to extend this work to stochastic simulator where the EI criterion (see Equation (2.13)) must take into account the heteroskedasticity of the emulator. When choosing sequential design for stochastic simulator, there is a trade-off between exploration and replication [16]. For the calibration purpose, this should be also done by integrating weights derived from the posterior distribution of the calibration parameters in their domain.

DoFE for calibration and validation. In most papers on calibration and validation, the DoFE is assumed to be given. However, if new experiments are affordable or if data collection is to be done while the simulator is already available, the framework of Bayesian experimental design [33] is of interest. The first steps are to translate the calibration and validation objectives into a utility function to be optimized to obtain the DoFE. Calibration and validation objectives are not equivalent, hence the utility function should be a trade-off of these objectives. Provided that several batches of field experiment are possible, another solution is to iteratively run batch experiments for alternatively enhancing the calibration precision and assessing the validity of the simulator.

The simulator may be actually valid only on some subregions of the input variable domain. Delimiting this domain is then of major interest and the associated question is to refine the DoFE and the DoNE to improve this delimitation. This could be done by using the mixture model proposed in [P3]. When the mixture model in Equation (2.2.3) gives posterior probabilities such that $\mathbb{P}(i_k \sim \mathcal{M}_1|\mathbf{y}^e) \approx \mathbb{P}(i_k \sim \mathcal{M}_0|\mathbf{y}^e) \approx 1/2$ (notation $i_k \sim \mathcal{M}_j$ ($j \in \{0,1\}$ means that the observation $y^e_{i_k}$ was generated under Model $\mathcal{M}_j$) for some observations $i_k \in \{1,\dots,n_e\}$, it means that there is too little information in the data to state or not the validity of the simulator in the neighborhoods of $\mathbf{x}^e_{i_k}$ s. Therefore running new field experiments in these neighborhoods could help to understand more precisely the departure of the simulator from the physical process.

Extrapolation. Furthermore, if the goal is to extrapolate from the simulator outside the domain of input variables $\mathbf{x}$ where no physical experiment is doable, the parametric uncertainty on the calibration parameter θ could be assessed from different choices of design. For instance, let us consider a ball drop experiment [41]. The behavior of a ball drop is modeled by the ODE:

$$\frac{d^2h}{dt^2} = g - \frac{C_d}{2} \cdot \frac{3\rho_{air}}{4\rho_{ball}R} \cdot \left(\frac{dh}{dt}\right)^2, \tag{2.32}$$

with initial condition $h(0) = h_0$, where h is the height of the ball, g is the gravity constant, C_d a coefficient for air resistance and ρ_{air}, ρ_{ball} the respective densities of air and of the ball. The scalar quantity of interest is the drop time i.e. the time t_d when

$h(t_d) = 0$. The simulator $f(x = h_0, \theta = (g, C_d, \rho_{air}, \rho_{ball})$ gives the drop time when the initial dropping height is h_0 and for a choice of physical parameters θ. The goal may be to use the simulator to predict the dropping time of a ball from a big height (let say $h_0 = 200m$) although the field experiments must be run from smaller heights with the constraint the higher, the more difficult. In this example, by doing a sensitivity analysis, we can show that only the parameter g matters if h_0 is small while when h_0 is large such as in the extrapolative setting, the three other parameters are important. Therefore, calibrating the parameters with a DoFE with too small heights will not reduce the parametric uncertainty on $(C_d, \rho_{air}, \rho_{ball})$ and the prediction uncertainty will be large. Then the goal is to find a trade-off between the number of experiments and the initial heights in the DoFE which enables a sufficient reduction of the prediction uncertainty in the extrapolative setting. Using a linear approximation in θ of the simulator

$$f(h_0, \theta) = \beta_0(h_0) + \sum_{j=1}^{p} \beta_j(h_0)\theta_j$$

may be a way to evaluate the reduction of parametric uncertainty provided by a particular DoFE.

3 Network Analyses

Résumé du chapitre en français

Un réseau permet de représenter des données d'interaction. Il correspond à un graphe constitué d'un ensemble de nœuds et d'arêtes indiquant quelles paires de nœuds sont en interaction. L'information d'interaction peut être plus riche qu'une distinction binaire entre interaction ou non. Le cas échéant, le réseau est dit valué si l'on a, par exemple, un comptage sur le nombre d'interactions observées ou une force d'interaction qui est quantifiée. Un cas particulier de réseau est un réseau bipartite qui représente les interactions entre deux ensembles de nœuds. Les interactions peuvent avoir lieu entre les nœuds des deux ensembles mais pas au sein des ensembles.

Dans mes contributions, j'ai considéré les réseaux dans trois contextes différents : i) le réseau est une des entrées d'un modèle complexe, le but étant de quantifier l'influence de sa topologie sur les sorties du modèle [JP12]; ii) le réseau est latent et il influence la dynamique d'une épidémie, le but étant alors d'inférer le réseau à partir d'observations de la dynamique [JP1]; iii) le réseau est observé, le but étant d'étudier sa structure en regroupant les nœuds ayant des profils de connexion similaires [JP8, JP11, JP13, JP5, P6, P2].

Le réseau a une importance primordiale dans des modèles de métapopulation avec extinction et colonisation qui sont équivalents à des modèles SIS (Susceptible Infecté Susceptible) en épidémiologie. Les colonisations ou les contaminations ne sont supposées possibles qu'au travers de contacts donnés par le réseau. Dans [JP12], nous avons étudié dans un modèle de métapopulation stochastique, l'influence de la topologie du réseau sur des indicateurs tels que la probabilité de persistance d'une population et l'occupation moyenne au bout d'un nombre fini de générations. Lorsque la persistance était menacée, nous avons mis en évidence que les topologies donnant lieu à des nœuds très connectés résistaient mieux que celles ayant des nœuds avec des connexions plus équilibrées tandis que ces dernières favorisaient une occupation moyenne plus importante lorsqu'il y a avait une grande probabilité de persistance.

Dans des modèles similaires type SIS, nous avons proposé dans [JP1] une inférence efficace du réseau de contact entre individus à partir des seules données de statuts des individus (infecté ou non) au cours du temps. Cette inférence du réseau est dérivée de probabilités calculées pour chaque arête d'avoir été source d'infection au moins une fois au cours de la fenêtre d'observation. Ces calculs s'appuient sur une utilisation du théorème arbre matrice qui permet de sommer efficacement sur tous les arbres de contamination possibles.

Les contributions concernant le regroupement des nœuds d'un réseau observé à partir des profils de connexion reposent sur des extensions des modèles à blocs stochastiques (SBM : Stochastic Block Model pour les réseaux simples et LBM : Latent Block

Model pour les réseaux bipartites). Ces modèles considèrent que l'hétérogénéité des connexions dépend de variables latentes associées aux nœuds. Un paramètre de connectivité est alors associé à chaque paire de variables latentes. Retrouver les variables latentes donne une classification des nœuds (leur appartenance aux différents blocs) et l'inférence des paramètres de connectivité permet de comprendre la structure du réseau. Nous nous sommes intéressés à différents types de réseaux multicouches à savoir multiplexes [JP8, JP11], multiniveaux [P2] et multipartites [P6]. Nous avons obtenu l'identifiabilité des extensions multiplexes et multiniveaux. Pour ces trois extensions, nous avons adapté l'inférence par algorithme variationnel espérance-maximisation (VEM) et la sélection du nombre de blocs latents par un critère de vraisemblance pénalisée. L'extension multiplexe a été motivée par des données en sociologie sur des interactions entre chercheurs pouvant être directes ou à travers leurs institutions. L'extension multiniveau est également motivée par des données en sociologie, concernant les interactions entre des commerciaux sur une foire de programmes télévisés pour lesquels on dispose également des contacts entre leurs entreprises. L'extension multipartite a été principalement motivée par des données en écologie pour pouvoir traiter conjointement des interactions entre des plantes et plusieurs types de visiteurs (pollinisateurs, fourmis, oiseaux) et des données en ethnobiologie où l'on s'intéresse à la circulation de semences entre fermiers et aux inventaires des variétés cultivées par ceux-ci. Dans [JP5], nous avons traité des données manquantes dans les réseaux d'interaction. Si ces données sont manquantes au hasard, adapter l'inférence est simple. Par contre, si ce n'est pas le cas, il est indispensable de prendre en compte la stratégie d'échantillonnage dans l'inférence afin d'obtenir des estimations non biaisées des modèles à blocs stochastiques.

Dans mes perspectives, je compte poursuivre mes travaux sur les inférences de réseau intervenant dans des modèles dynamiques. L'inférence pourra se limiter à des caractéristiques résumées du réseau si les observations sont trop partielles. Elle pourra aussi intégrer des données plus complexes telles que des données génétiques pour les différentes sous-populations d'une métapopulation. L'étude de l'inférence des modèles à blocs stochastiques à partir de réseaux partiellement observés pourra également être poursuivie notamment dans le cadre des réseaux écologiques d'interaction. En effet, le recueil des données présente un cadre quelque peu différent où les interactions observées sont fiables mais une incertitude pèse sur les interactions non observées. Cette absence d'observation peut résulter d'un effort d'échantillonnage trop faible plutôt que d'une cause biologique. Grâce à des mesures de la complétude d'échantillonnage, la modélisation du processus d'observation pourrait corriger l'inférence, ceci pour des données provenant de campagnes scientifiques mais aussi pour des données issues des sciences participatives. Également en écologie, faire le lien entre un indicateur tel que la robustesse qui mesure la capacité du réseau à supporter des perturbations et la structure du modèle en blocs stochastiques est une perspective intéressante puisqu'elle permettrait de compenser des effets d'échantillonnage et faciliterait des comparaisons de robustesse entre différents réseaux échantillonnés dans des conditions différentes.

In my work, I have considered networks in three distinct contexts: i) a network is an input of a complex model and the goal is to assess to what extent the topology of the network impacts the outputs [JP12]; ii) the network is not observed but latent, the goal is to infer it from data whose conditional dependencies rely on the network [JP1]; iii) the network is observed and the goal is to unravel its structure by clustering its node according to their connections [JP8, JP11, JP13, JP5, P6, P2]. Before presenting the three contexts and the related contributions, we provide the common notations and terminology.

3.1 Notations and Terminology

A network is a way to represent interaction data. It corresponds to the mathematical object named graph. A network is then given as a collection of two sets: a set of nodes $\mathcal{N}$ (a.k.a. vertices) generally numbered from 1 to n (i.e. $\mathcal{N} = \{1,\dots,n\}$) and a set of edges (a.k.a. links, ties, connections) $\mathcal{E}$ representing the subset of pairs of nodes (a.k.a dyads) which are linked. The nodes are the individuals which may be in interaction and the edges are the interactions which do exist. We have the inclusion $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$. The edges may be directed/oriented or not depending on the reciprocity of the interaction. For a directed network, we will use the convention $(i, j) \in \mathcal{E}$ means that there is a link from i to j. Moreover, an edge may carry more information than simply a binary existence / non-existence of an interaction. It may also contain information on the strength or frequency of this interaction. In this case, besides the set of edges $\mathcal{E}$, a corresponding set of values is provided. Such a network is called a valued network; otherwise a binary network.

A common representation for a network is as an adjacency matrix denoted by A as illustrated in Figure 3.1 for a binary network. If $(i, j) \in \mathcal{E}$, $A_{ij} = 1$, otherwise $A_{ij} = 0$. Usually, the network does not have any self-loop (a node is not connected to itself) then the diagonal is a null vector. The adjacency matrix may be either symmetric if the network is undirected or non-symmetric otherwise. If the network is valued, the matrix A may contain the strength of interaction between nodes instead of 0/1. A particular kind of network is bipartite network when there are two sets of nodes which may be in interaction between sets but no interaction occurs within a set. They are particularly relevant in ecology for representing plant-pollinator interactions for example. In this case, the network is rather represented by an incidence matrix which is rectangular and has n rows (size of the first set of nodes; plants e.g.) and m columns (size of the second set of nodes; pollinators e.g.). Figure 3.2 displays an example of a bipartite network and the corresponding incidence matrix. The adjacency matrix of a bipartite network is recovered from the incidence matrix as $A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$.

A network is often summarized by statistics which are computed at the node level or at the full network level [96]. A classical statistic at the node level is the degree that is to say the number of edges involving a given node. It is computed as $d_i = \sum_j A_{ij}$ if the network is undirected, otherwise indegree and outdegree are to be defined.

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

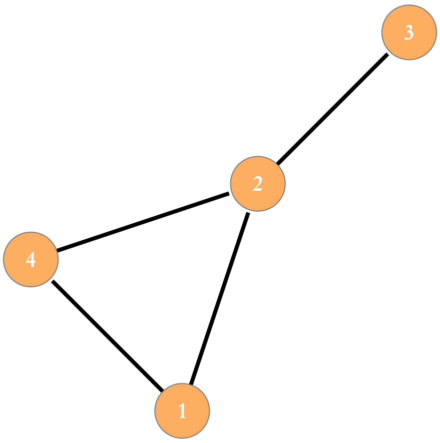


Figure 3.1 – *Adjacency matrix and corresponding representation of an undirected binary network.*

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

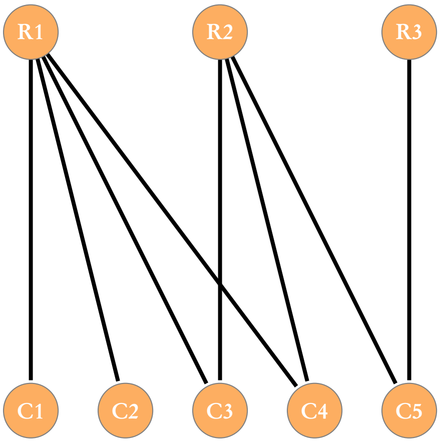


Figure 3.2 – *Incidence matrix and corresponding representation of a bipartite binary network. R (respectively C) stands for the nodes corresponding to the rows (respectively columns) of the incidence matrix B.*

3.2 Influence of the Network in Complex Processes

Dynamic extinction colonization models (also called contact processes) are widely studied in epidemiology and in metapopulation theory. Contacts are usually assumed to be possible only through a network of connected patches. This network accounts for a spatial landscape or a social organization of interactions. To study the persistence of a metapopulation in such a model, several papers have used deterministic models where the evolution is described by differential equations [110, 82, 158]. These models are grounded on an asymptotic approximation in the number of patches. The same models are used in epidemiology (SIS: Susceptible Infected Susceptible model). More recently, some studies have dealt with the stochastic effect due to a finite and limited number of patches/actors. Chakrabarti el al. [32] have proposed an approximation in the stochastic model which leads to conclusions similar to the ones obtained with deterministic models. Gilarranz and Bascompte [69] have shown by simulations the impact of stochasticity due to a limited number of patches and they have underscored the differences with the results obtained with deterministic models when comparing the ability of different networks to conserve a metapopulation. However, their results depend only on the ratio of the extinction rate to the colonization rate which is not relevant in a stochastic model. Our contribution [JP12] was motivated by the characterization of a seed circulation network among farmers. This question has arisen in the study of the Réseau Semences Paysannes, an emergent French farmers' organisation.

Dynamic extinction-colonization model. The dynamic model under study describes the presence or absence of a crop variety on n different farms (patches according to metapopulation vocabulary) during a discrete time evolution process. This metapopulation is identified with an undirected network G with n nodes (patches or farms) and adjacency matrix $A = [A_{ij}]_{i,j}$ where $A_{ij} = 1$ if patches i and j are connected ($i \sim j$) and 0 otherwise. We further denote by Z_i^t the occupancy of patch i ($i = 1 \dots n$) at time t, namely $Z_i^t = 1$ if patch i is occupied at time t and 0 otherwise. The vector $Z^t = [Z_i^t]_i$ depicts the composition of the whole metapopulation at time t. A time step corresponds to a generation of culture. Between two generations, two events may occur: extinction and colonization with respective rates e and c. Within each time step, extinction events first take place and occur in occupied patches independently of the others, with a probability e, supposed to be constant over patches and time. Colonizations events then take place and are only possible between patches linked according to the static relational network G. An empty patch may be colonized by an occupied patch with a probability c. This probability is also assumed constant over linked patches and time steps. Thus, the probability that the patch i, if empty at generation t, is colonized between generations t and $t + 1$ is:

$$\mathbb{P}(Z_i^t = 1 | Z_i^{t-1} = 0; (Z_j^{t-1})_{j \neq i}) = 1 - (1 - c)^{o_{i,t-1}} \tag{3.1}$$

where $o_{i,t} = \sum_j a_{ij} Z_j^t$ is the number of its occupied neighbours at generation t. This model is similar to the one proposed in [69] and also to the epidemic model SIS used in [32] where the nodes are the individuals, the two possible states for an individual are susceptible or infected and the network depicts the potential contact between individuals.

The stochastic process $(Z_t)_{t \in \mathbb{N}}$ is a discrete time Markov chain with 2^n possible states. As given in [51], the matrices describing the colonization C and the extinction

E can be computed and the transition matrix of $(Z_t)_{t\in\mathbb{N}}$ is obtained as the product of these two matrices: $M = E\cdot C$. This Markov chain is irreducible and aperiodic provided that the adjacency matrix A of the social network has only one connected component. If $e > 0$, there is a unique stationary distribution which is the absorbing state where all patches are empty meaning that the crop is extinct. The extinction time $T_0 = \inf\{t > 0, \#Z_t = 0\}$, where $\#Z_t = \sum_{i=1}^{n} Z_i^t$ is the number of occupied patches at time t, is such that $\mathbb{P}(T_0 < \infty) = 1$. Contrary to a deterministic model, we cannot separate parameter settings which lead to extinction from the others. Therefore, we choose a number of generations T and we consider the following indicators to characterize the persistence: the probability of persistence $\mathbb{P}(T_0 > T)$ at generation T and the mean number of occupied patches $\mathbb{E}(\#Z_T)$ at generation T. They can be computed by exact computation from the transition matrix provided that n is not too large ($n \sim< 10$). Otherwise, several simulations are run to estimate them.

Analysis of the dynamic model. Put in the context of UQ, the dynamic extinction colonization model is a stochastic simulator which depends on parameters e, c and G. We want to conduct a sensitivity analysis with respect to two specific scalar outputs: $\mathbb{P}(T_0 > T)$ and $\mathbb{E}(\#Z_T)$. Since the input G is the contact network, it needs to be characterized by scalar or categorical variables. We choose to describe the network by its density or equivalently by its number of edges and its topology which is the way of distributing edges among nodes. We only consider networks with the constraint of a unique connected component. As illustrated in Figure 3.3, five contrasted topologies are compared: i) Erdős-Rényi [61] topology where the edges are drawn uniformly and independently among the set of possible edges, ii) community network simulated under an Stochastic Block Model (SBM) [131] with equal size communities having a larger probability of connection within a community than between communities, iii) lattice network where the degree of the nodes are chosen as homogeneous as possible, iv) and v) a preferential attachment topology [2] leading to a high heterogeneity of degrees. In the preferential attachment topology, the nodes are added sequentially. At each step, a single node is added and is connected to the nodes already in the network with probability

$$\mathbb{P}(\text{connection to node } k) \propto d_k^b ,$$

where the power b is chosen in order to tune the strength of the preferential attachment.

The sensitivity analysis of the dynamic extinction colonization model is performed through an analysis of variance model with respect to the four inputs: e, c, the number of edges n_e and the topology. The three first parameters are discretized in three levels such that we explore a diversity of situations ranging from likely persistence to likely extinction in T generations. The topology is chosen among the five ones described above. The comparison based on $\mathbb{E}(\#Z_{100})$ has shown an inversion in the ranking of the topologies which was similar to the one noticed by [69]. On the one hand, when the combination of values of e, c and d ensured persistence with a high probability, the best topologies were those with a better balance in degree distribution such as the lattice, ER and community topologies. However, although the difference in mean was found significant, the order of magnitude of this difference was only of a few patches (≈ 5) for $n = 100$ patches. On the other hand, the topologies leading to some very connected (hub) such as the preferential attachment topologies (especially when the power parameter is set at 3) maximized the number of occupied patches when the persistence

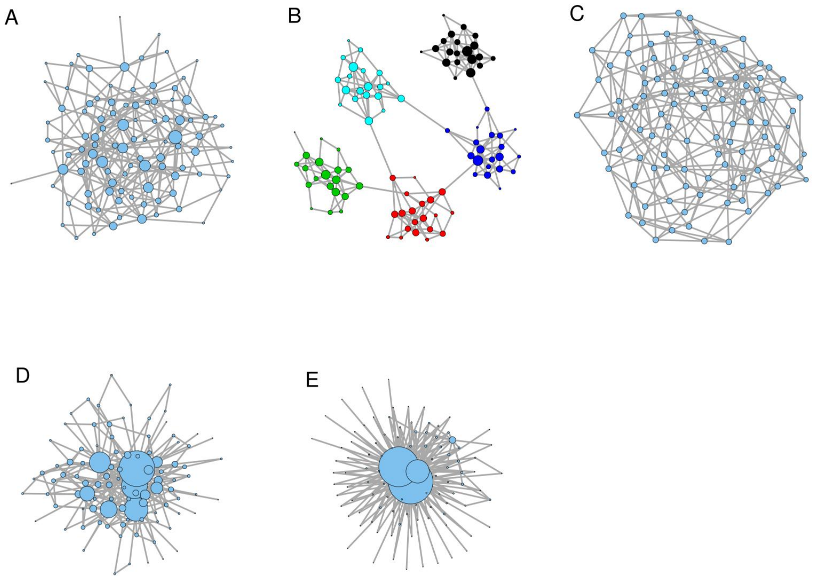


Figure 3.3 – *Simulation of networks with* 100 *nodes and* 247 *edges according to Erdős-Rényi model (**A**), community model (**B**), lattice model (**C**), preferential attachment model with power 1 (**D**) and power 3 (**E**). The size of a node is proportional to its degree.*

in the system is threatened in 100 generations. In that case, the differences were more contrasted between topologies as illustrated in Figure 3.4. Different scenarios in the context of the Réseau Semences Paysannes are also explored in [JP12].

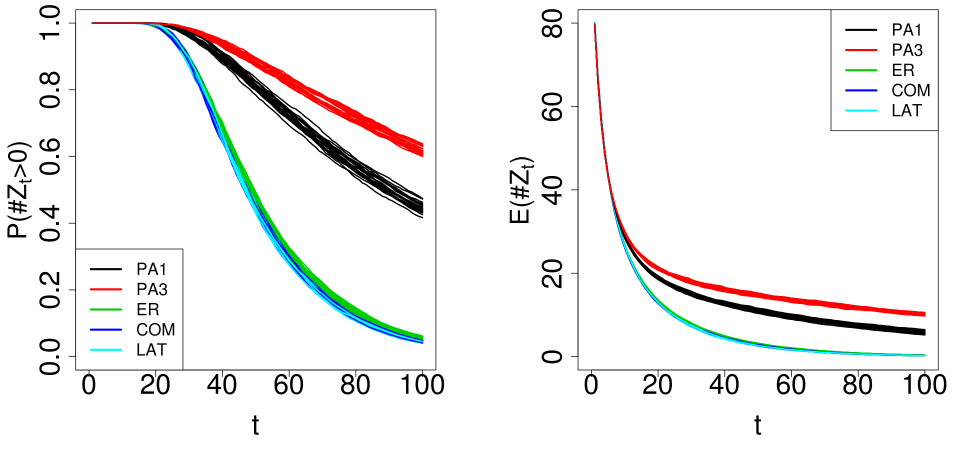


Figure 3.4 – *Probability of persistence (left) and mean number of occupied patches (right), in varying t generations (based on* 20 *replications of the network for a given topology) for $n = 100$, $c = 0.01$, $e = 0.25$ and $n_e = 30\% \times n(n-1)/2$. COM: community network, ER: Erdős-Rényi network, LAT: Lattice network, PA1: preferential attachment network with power 1, PA3: preferential attachment with power 3.*

3.3 Inference of Networks

Assuming a similar dynamic model as in [JP12] presented in the previous section, the question could be to infer the contact network from the binary status of individuals observed throughout time (occupied / empty in a metapopulation context or susceptible / infected in an epidemic context). We proposed in [JP1] to compute the probability for each edge to be part of the contact network by using the matrix tree theorem on the set of vertices made of the individual status at all times. This leads to a cheap computational complexity of order $\mathcal{O}(mn^2)$, where n is the number of nodes and m the length of the time series. The efficiency is demonstrated on synthetic examples and two applications on real datasets concerned with seed choices by farmers in India and a measles outbreak are dealt with in the paper.

Dynamic model. The dynamic model here has the same extinction (or curation) process as the one defined in 3.2, independent extinction with same probability e along time. A first slight difference is that, in this model, colonization only happens if the node was empty at the previous generation. In the model defined in Section 3.2, an extinction may be immediately followed by a colonization between two consecutive time steps. Second, although the colonization (or infection) process is still independent conditionally on the previous time step, the probability of colonization is different from Equation (3.1) and is here given by:

$$\mathbb{P}(Z_i^{t+1} = 1 | Z_i^t = 0; Z_{pa(i,t)}^t = 1) = c \tag{3.2}$$

where c is a colonization probability and $pa(i,t)$ is the node which is the parent of node i from time t to time $t+1$. Between two consecutive time steps, each node may be colonized by a unique node which is called its parent. This parent node may change from one step to another. Therefore, the colonization path is actually a tree, once a root vertex Δ and edges from Δ to nodes $(i,1), 1 \leqslant i \leqslant n$, have been added (*cf.* Figure 3.5 for an illustration). Formally, we let $\mathcal{N}^* := \{\Delta\} \cup \mathcal{N}$ be the augmented set with $nm+1$ elements and we let T denote the tree on $\mathcal{N}^*$ resulting from the completion of the colonization path with the root Δ

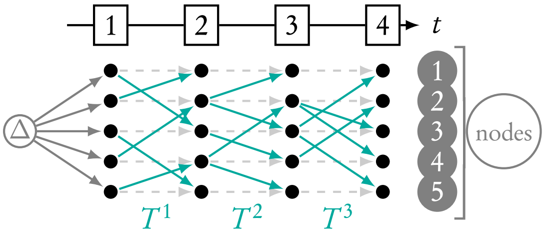


Figure 3.5 – *Graphical model associated to an example of tree $T = (T^1, T^2, \ldots)$.*

Because its edges only link vertices at time t to vertices at time $t+1$, T can be sliced into $m-1$ oriented bipartite graphs T^t $(1 \leq t \leq m-1)$, that each defines the parents nodes for the transition from time t to time $t+1$. More specifically, we denote by $\{[ij] \in T^t\}$ the event that makes i the parent of j during the transition from time t to time $t+1$. In the proposed modeling, the tree T is random and its distribution is defined as follows. We associate a prior weight β_{ij} with each oriented edge $[ij]$ and assume that, at each time t, each node j samples its parent i with probability proportional to β_{ij}. As a consequence, the probability of a tree T is

$$\pi(T) = B^{-1} \prod_{t=1}^{m-1} \prod_{[ij]\in T^t} \beta_{ij} \tag{3.3}$$

where $B := \sum_T \prod_{t=1}^{m-1} \prod_{[ij]\in T^t} \beta_{ij}$. The weights β_{ij} can be seen as a way to account for some prior knowledge about the likelihood of each edge or as parameters of the model that need to be inferred.

Then, conditionally on the tree structure the distribution of the data is derived by using the Markov assumption:

$$\pi(Z \mid T) = \pi(Z^1) \prod_{t=1}^{m-1} \pi(Z^{t+1} \mid Z^t, T^t) = \prod_{j=1}^{n} \pi(Z_j^1) \prod_{t=1}^{m-1} \prod_{i,j:[ij]\in T^t} \phi_{ij}^t,$$

where $\phi_{ij}^t := \pi(Z_j^{t+1} \mid Z_j^t, Z_i^t, [ij] \in T^t)$ which is given by

$$\begin{array}{cc|cc} \multicolumn{2}{c|}{\phi_{ij}^t} & Z_j^{t+1} = 1 & Z_j^{t+1} = 0 \\ \hline Z_j^t = 1 & Z_i^t = 1 & 1-e & e \\ Z_j^t = 1 & Z_i^t = 0 & 1-e & e \\ Z_j^t = 0 & Z_i^t = 1 & c & 1-c \\ Z_j^t = 0 & Z_i^t = 0 & 0 & 1 \end{array} . \tag{3.4}$$

Remark. In the proposed modeling the marginal probability for a susceptible node to get colonized depends on the fraction of colonized nodes in an implicit manner, through the choice of its parent (which may or may not be colonized). A more explicit dependence, such as the one given by Equation (3.1) cannot be cast in the tree-structured model we propose as it introduces a dependence with respect to the whole population.

Inferring the contact network. Taking advantage of the tree structure of the dynamic contact process T, we use the matrix tree theorem [31] to compute the sum over the latent tree set as a cofactor of the Laplacian matrix associated to the matrix containing the terms $\psi_{ij}^t = \phi_{ij}^t \beta_{ij}$. Therefore, the likelihood deriving from the probability distribution of Z is:

$$\ell(e, c; Z) = \pi(Z) = \sum_T \pi(Z \mid T)\pi(T) = \prod_{t=1}^{m-1} \prod_j \psi_{+j}^t \Big/ \prod_j (\beta_{+j})^{m-1}$$

where $\beta_{+j} := \sum_i \beta_{ij}$ and $\psi_{+j}^t := \sum_i \psi_{ij}^t$, that is

$$\psi_{+j}^t = \begin{cases} (1-e)\sum_i \beta_{ij} & \text{if } Z_j^t = 1, \; Z_j^{t+1} = 1, \\ e \sum_i \beta_{ij} & \text{if } Z_j^t = 1, \; Z_j^{t+1} = 0, \\ c \sum_i \beta_{ij} Z_i^t & \text{if } Z_j^t = 0, \; Z_j^{t+1} = 1, \\ -c \sum_i \beta_{ij} Z_i^t + \sum_i \beta_{ij} & \text{if } Z_j^t = 0, \; Z_j^{t+1} = 0. \end{cases}$$

From this likelihood expression, we can derive maximum likelihood estimates (MLE) for e and c. The estimate for e is explicit while the one for c corresponds

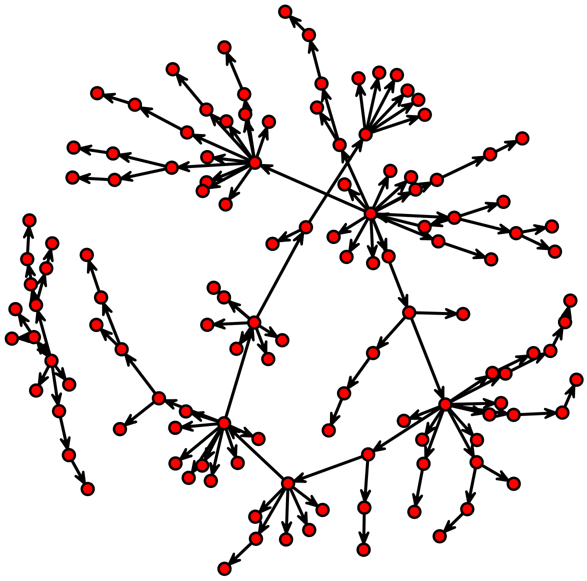


Figure 3.6 – *Network of seed choice influences between farmers. This network was obtained by taking the most probable neighbor for each farmer from the inferred edge probabilities.*

to an optimization. If we plug the MLE of e and c, we compute the probability that an edge $[ij]$ was used at least once in the dynamic evolution. More precisely, we consider the complementary sets on trees $\mathscr{T}_{ij} := \{T \in \mathscr{T} : \exists t, [ij] \in T^t\}$ and $\overline{\mathscr{T}}_{ij} := \{T \in \mathscr{T} : \forall t, [ij] \notin T^t\}$ and the corresponding events $E_{ij} := \{T \in \mathscr{T}_{ij}\}$ and $\overline{E_{ij}} := \{T \in \overline{\mathscr{T}}_{ij}\}$, where the latter states that $[ij]$ never appears along the tree T. To assess whether the edge $[ij]$ is part of the network, we want to compute the conditional probability

$$\mathbb{P}(\mathbb{E}_{ij} \mid Z) = 1 - \mathbb{P}(\overline{E_{ij}} \mid Z) = 1 - \mathbb{P}(\overline{E_{ij}}, Z)/\pi(Z) = 1 - \prod_t \left(1 - \frac{\psi^t_{ij}}{\psi^t_{+j}}\right). \qquad (3.5)$$

Remark. Other estimation methods can be used. In a Bayesian approach, the parameters β are seen as a way to set a prior on trees. The posterior distribution of parameters e and c can be established easily when using conjugate priors. The posterior distribution of edge probabilities can then be obtained as in Equation (3.5) via Monte Carlo sampling from the posterior distribution of parameters. Another approach is to see the proposed model as a mixture model with as many components as possible trees. In this setting, the weights β_{ij} act as the parameters ruling the proportions of the mixture components and their maximum likelihood estimates can be obtained via the EM algorithm [55]. One interest of this approach is that it allows the estimation the weights of the edges β_{ij}, rather than keeping them fixed at a prescribed value. In practice, none of these alternatives turned out to significantly improve edge retrieval.

Results. The performance of our method was assessed on synthetic data. A simulated network linking the n nodes was drawn from a given topology (Erdős-Rényi or preferential attachment). The components of the tree T^t giving the parent of the nodes at a given time step were enforced to have their edges among the edges of the simulated network. Then, the dynamic model was run from a unique node occupied and all the others empty. Finally, the accuracy of our edge inference method was assessed by measuring its ability to recover the edge of the simulated network. More precisely, we computed area under curve (AUC) by comparing the conditional edge probabilities given in Equation (3.5) to the actual edges of this network. A global result is that all the AUC are larger than 0.5 meaning that it is possible to recover the edges based only on the observation of nodes status along time. Moreover, the edge inference is easier when the parameter c is not too large, the network is not too dense and when its topology is ER. It is indeed easier to learn from the early time steps of the propagation if the colonization is not to fast. We then applied our method on real datasets. In collaboration with Andrew Flachs and Glenn Stone, we used data on seed inventory along several years in India [63]. We inferred a network of farmers which aims to capture the influence among them in seed choice. It is represented in Figure 3.6 by selecting the most probable edge for each farmer. We mainly noticed that some nodes seem to have a great influence while some others are organised in small groups. By using covariates on farmers, we found that edges within the same caste and the same village were more likely to occur than between.

3.4 Analysis of Networks

A network may be analyzed through the computation of descriptive statistics [96]. Depending on the scientific community, some specific statistics are focused on; reciprocity, transition [17] and an emerging power law distribution of the degrees in social sciences [127]; modularity and nestedness [128, 167, 64] in ecology. To assess the significance of a computed value of the statistic, it is compared to its distribution under a null model. This distribution is obtained by resampling the network under some constraints (on degrees for instance, see [168] in ecology). Another way to analyze networks is to resort to probabilistic generative models. They are powerful tools to model the heterogeneity of connections in networks and they have the advantage of being agnostic since they do not look for a particular property of the network. When a probabilistic model is fitted, a particular structure, such as modularity, transitivity, nestedness, is not sought for. The goal is to unravel the structure of the network from the fitted component of the probabilistic model.

The simplest random graph model for a network is the Erdős-Rényi model [61] where all dyads are independent and the probability of an edge is the same. A possible extension is the Exponential Random Graph Model (ERGM) [143] which is popular in social science. In this model, the distribution of the graph belongs to the exponential family and the sufficient statistics count for some local motifs in the network such as the number of edges, of triangles. In my contributions, I do not consider this model but I rather focus on latent variable models. The latent variable models for graph are numerous (see [122] for a recent review). In these models, a latent variable is associated with each node and the connectivity of a node depends on its latent variable. The latent variable may lie in a continuous space (Latent Position Cluster Model [81] or Latent Position Model [34]) or in a discrete space (Stochastic Block Model, SBM [156]). In my contributions, I worked with the SBM which is flexible enough to be extended quite naturally to handle multilayer networks. Moreover, the SBM provides a clustering of all the nodes in the network by recovering the latent discrete variables which is an interesting feature to analyze the structure of the network.

Note that in this section, we will use the notation $\mathbf{Y}$ to denote the random variables which are defining the network. They are indexed over the set of dyads. Thus, the adjacency matrix is the modeled stochastic object and that is why we use a different notation from the previous notation A.

3.4.1 Background on Block Models for Networks

3.4.1.1 Block Models

The SBM consists in a mixture model on the dyads of a simple network. A random variable is associated with each dyad, the distribution of this random variable depends on the latent variables associate with the two nodes involved in the dyad. More specifically, we denote by $\mathbf{Z} = (Z_1, \ldots, Z_n)$ the latent variables which give the clusters / blocks the nodes belong to. These latent variables are in the set $\{1, \ldots, K\}$ where K is the number of blocks. For all $i = 1, \ldots, n$, we assume

$$\mathbb{P}(Z_i = k) \overset{ind}{=} \pi_k.$$

Conditionally on $\mathbf{Z}$, the random variables on dyads are denoted by Y_{ij} and they follow independently:

$$Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathscr{F}(\alpha_{Z_i, Z_j}). \tag{3.6}$$

The random variables Y_{ij} are indexed by the set of dyads denoted by $\mathscr{A}$. This set depends on the network, we may have $\mathscr{A} = \{1, \ldots, n\}^2$, $\mathscr{A} = \{(i,j) : 1 \leq i, j \leq n, i \neq j\}$ if there is no loop or $\mathscr{A} = \{(i,j) : 1 \leq i < j \leq n\}$ if there is no loop and the relation is not oriented. Depending on the nature of the links, the distribution $\mathscr{F}$ may be a Bernoulli distribution if the relations are binary (interaction or not), a Poisson distribution if they correspond to a counting (number of interactions), a Gaussian distribution (continuous measure of the strength of interaction). The distribution $\mathscr{F}$ may be given with respect to some covariates which are either at the node level $(X_i)_{1 \leq i \leq n}$ or at the dyad level $(X_{ij})_{(i,j) \in \mathscr{A}}$. In the former case, the covariates may be either transferred at the dyad level by using a function ϕ computing a distance betweeen them, for instance: $X_{ij} = \phi(X_i, X_j) = ||X_i - X_j||$, or incorporated as two covariates: X_i and X_j for the dyad (i,j). When the covariates are transferred at the dyad level the distribution is then: $Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathscr{F}(\alpha_{Z_i, Z_j} + \beta X_{ij})$. Otherwise, the distribution is: $Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathscr{F}(\alpha_{Z_i, Z_j} + \beta_1 X_i + \beta_2 X_j)$. The latter expression makes sense provided that the relation is oriented. Although most of the literature on SBM focuses on binary networks [131, 50], modeling valued networks with an SBM is rather natural [115] and the SBM can be even used when the observations on the dyads are more complex such as longitudinal observations [121] or text data [20]. The SBM for binary network has been proven to be identifiable in [30].

By assuming that the nodes belong to blocks that shape their connectivity profile, the SBM captures the heterogeneity of connections that does exist in many real networks. The SBM encompasses a wide variety of typical structure of interaction. Figure 3.7 provides an illustration of three particular binary network structures modeled by an SBM. The matrix α contains this structure. We display a network visualization with the nodes being colored according to their block belonging and the reordered adjacency matrix. The network visualization is not always a simple task and may fail to unravel the structure. The reordering of the adjacency matrix is sometimes more informative. Note that on real dataset the blocks are unknown since the variables in $\mathbf{Z}$ are latent. Therefore, neither the coloring of nodes nor the reordering of the adjacency matrix is possible until the blocks are recovered by the inference procedure. The assortative structure is maybe the first that comes to mind, the blocks are communities where the connections are more likely within than between. This structure is similar to modularity [128]. The nested structure is also a common structure which is looked for, especially in ecology [64]. The blocks may be organized from the most central / generalized one to the less central one. In this structure, the individuals belonging to the least connected block are more likely to be connected with the individuals belonging to the most connected blocks. Finally, a bipartite like structure is displayed where two groups of two blocks are paired since they are interacting mainly with each other.

We presented above the SBM for simple network. When the network is really bipartite the adapted corresponding model is the Latent Block Model (LBM) [74, 94] a.k.a. bipartite SBM (biSBM) [104]. In this case, there are two sets of latent variables $\mathbf{Z}^1 = (Z^1_1, \dots, Z^1_n) \in \{1, \dots, K_1\}^n$ and $\mathbf{Z}^2 = (Z^2_1, \dots, Z^2_m) \in \{1, \dots, K_2\}^m$. All these latent variables are independent and their distributions are given by the vectors of parameters π^1 and π^2: $\mathbb{P}(Z^1_i = k) = \pi^1_k$ for all $i \in \{1, \dots, n\}$, $k \in \{1, \dots, K_1\}$ and $\mathbb{P}(Z^2_j = l) = \pi^2_l$ for all $j \in \{1, \dots, m\}$, $l \in \{1, \dots, K_2\}$. These latent variables shape the heterogeneity of interactions: for all $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$Y_{ij} | Z^1_i, Z^2_j \overset{ind}{\sim} \mathcal{F}(\alpha_{Z^1_i, Z^2_j}). \tag{3.7}$$

The matrix α is then rectangular of dimension $K_1 \times K_2$.

A limitation of the SBM is that the expected degrees are the same for all nodes in the same block. Two extensions have been proposed to introduce more heterogeneity in the degree distribution, Degree-Corrected SBM (DCSBM) [91] and Popularity-Adjusted Block Model (PABM) [154, 130]. In the DCSBM, the distribution of Y_{ij} is either Bernoulli or Poisson and given by: $Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathcal{F}(\lambda_i \lambda_j \alpha_{Z_i, Z_j})$ where λ is a vector with n elements with some constraints to ensure identifiability. This parameter introduces more diversity among the degrees. In a classical SBM, the matrix α may model both the degree diversity and the specific connection preferences whereas it models only the connection preferences in a DCSBM. Therefore, the clustering obtained with a DCSBM is not based on the differences of degrees. The PABM considers that $Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathcal{F}(\lambda_{iZ_j} \lambda_{jZ_i})$ where the matrix $\Lambda = (\lambda)_{1 \leq i \leq n, 1 \leq q \leq K}$ must satisfies some identifiability constraints. This matrix gives the popularity of each node with respect to each block / community. The PABM is a generalization of DCSBM and SBM [154]. Its flexibility lies in the tuning of the degrees respectively to the blocks. In the DCSBM, a large λ_i increases uniformly the probability of connection of node i with any other nodes no matter their respective blocks whereas in PABM, K parameters $(\lambda_{i1}, \dots, \lambda_{iK})$ tune the probabilities of connection of node i with nodes from the other blocks.

Another direction to generalize the SBM is to have a more complex latent structure by allowing either the nodes to belong to several latent blocks which is the overlapping SBM (OSBM) [105] or by making each node drawing its block for each dyad it is involved in which is the Mixed Membership SBM (MMSBM) [1]. In the OSBM, for each node i the vectors $\mathbf{Z}_i \in \{0, 1\}^K$ are drawn as $\mathbf{Z}_i \overset{ind}{\sim} \prod_{i=1}^K \text{Bern}(\pi_k)$ and then $Y_{ij} | \mathbf{Z}_i, \mathbf{Z}_j \overset{ind}{\sim} \mathcal{F}(\mathbf{Z}_i^T W \mathbf{Z}_j + \mathbf{Z}_i^T U + \mathbf{Z}_j^T V + w^*)$ where W is a $K \times K$ matrix accounting for the interaction between the blocks, U, V are K vectors accounting for specific outgoing and ongoing effects and w^* a scalar acting as an offset. The MMSBM assumes that

$$\alpha = \begin{pmatrix} 0.70 & 0.09 & 0.09 \\ 0.09 & 0.70 & 0.09 \\ 0.09 & 0.09 & 0.70 \end{pmatrix}$$

$$\alpha = \begin{pmatrix} 0.70 & 0.70 & 0.70 & 0.70 \\ 0.70 & 0.70 & 0.70 & 0.09 \\ 0.70 & 0.70 & 0.09 & 0.09 \\ 0.70 & 0.09 & 0.09 & 0.09 \end{pmatrix}$$

$$\alpha = \begin{pmatrix} 0.09 & 0.70 & 0.09 & 0.09 \\ 0.70 & 0.09 & 0.09 & 0.09 \\ 0.09 & 0.09 & 0.09 & 0.70 \\ 0.09 & 0.09 & 0.70 & 0.09 \end{pmatrix}$$

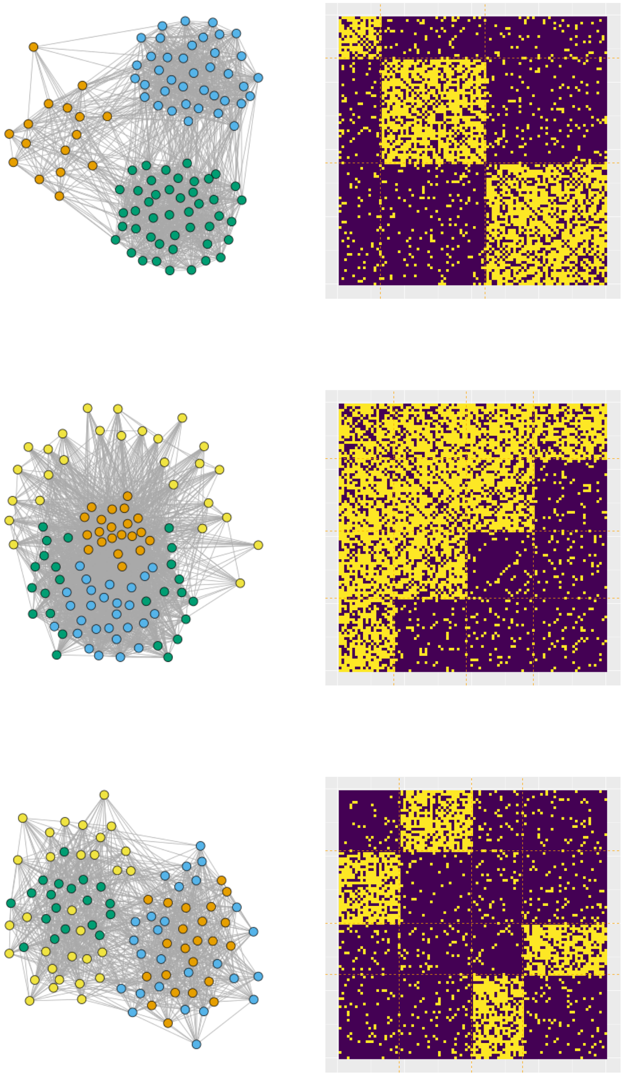


Figure 3.7 – *Three particular binary network structures modeled through an SBM. First row corresponds to an assortative structure, second row to a nested structure, third row to a bipartite like structure. Left column gives the matrices of $\alpha = \mathbb{P}(Y_{ij} = 1|Z_i, Z_j)$, middle column is a network plot where the colors correspond to the different blocks, right column is the reordered adjacency matrix with yellow square representing actual edges ($Y_{ij} = 1$) and purple absence of edge.*

some weights are drawn for each node $\gamma_i \overset{ind}{\sim} \text{Dirichlet}(\pi_1, \dots, \pi_K)$. Then for each dyad (i,j), two latent variables $Z_{i\to j}$ and $Z_{j\to i}$ are drawn independently with probabilities: $\forall (k,l) \in \{1,\dots,K\}^2$, $\mathbb{P}(Z_{i\to j}) = \gamma_{ik}$ and $\mathbb{P}(Z_{j\to i}) = \gamma_{jk}$. They impact the distribution of $Y_{ij} \overset{ind}{\sim} \mathscr{F}(\alpha_{Z_{i\to j}Z_{j\to i}})$. The mixed membership then lies in the distribution of weights.

3.4.1.2 Inference and Model Selection

When we are provided with a network $\mathbf{Y} = (Y_{i,j})_{(i,j)\in\mathscr{A}}$, the goal is to estimate the parameter of the SBM as well as recover the latent variables. In addition to the inference task, the number of blocks K is unknown and shall be chosen according to the data.

Variational EM algorithm. We start by detailing a variational EM (VEM) inference procedure for a known K. Since the SBM is a latent variable model, the EM algorithm [55] appears as a natural idea for inferring it. However, its particular dependency structure makes the EM algorithm intractable. Thus, a variational approach was proposed in [50].

We denote the parameters by $\theta = (\alpha, \pi)$. Then, the complete likelihood writes as

$$\log \ell_c(\theta; \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}_{\{Z_i=k\}}\pi_k + \sum_{(i,i')\in\mathscr{A}} \sum_{(k,k')=1}^{K} \mathbb{1}_{\{Z_i=k, Z_{i'}=k'\}} f(Y_{ii'}, \alpha_{kk'}) \quad (3.8)$$

where f is the log-density of $\mathscr{F}$. Thus the observed likelihood is:

$$\ell(\theta = (\alpha,\pi); \mathbf{Y}) = \sum_{\mathbf{Z}\in\{1,\dots,K\}^n} \log \ell_c(\theta; \mathbf{Y}, \mathbf{Z}). \quad (3.9)$$

The sum over Z is then intractable as soon as either n or K becomes large. The EM algorithm is not practicable here since the distribution of $\mathbf{Z}$ conditioned to $\mathbf{Y}$ ($\mathbb{P}(\cdot|\mathbf{Y};\theta)$) is not tractable because of the dependencies in $\mathbf{Z}$. In lieu of this posterior distribution, we use a variational approximation which consists of seeking for a distribution such that the elements of $\mathbf{Z}$ are independent:

$$\mathscr{R}_{\mathbf{Y},\tau}(\mathbf{Z}) = \prod_{i=1}^{n}\prod_{k=1}^{K} (\tau_{ik})^{\mathbb{1}_{Z_i=k}}, \quad \text{where} \quad \tau_{ik} = \mathbb{P}_{\mathscr{R}_{\mathbf{Y},\tau}}(Z_i = k). \quad (3.10)$$

The goal is then to maximize the lower bound of the likelihood with respect to θ and $\mathscr{R}$:

$$\mathscr{I}_\theta(\mathscr{R}) = \log \ell(\theta; \mathbf{Y}) - \text{KL}[\mathscr{R}, \mathbb{P}(\cdot|\mathbf{Y};\theta)], \quad (3.11)$$

where KL stands for the Kullback-Leibler divergence.

The VEM algorithm produces a sequence $(\mathscr{R}_{\mathbf{Y},\tau^{(t)}}, \theta^{(t)})$ by alternating the two steps:

VE step: Find $\tau^{(t+1)}$ by maximizing $\mathscr{I}_{\theta^{(t)}}(\mathscr{R}_{\mathbf{Y},\tau})$. This update is not explicit and can be achieved through fixed point relations.

M step: Find $\theta^{(t+1)}$ by maximizing $\mathscr{I}_\theta(\mathscr{R}_{\mathbf{Y},\tau^{(t+1)}})$ which is equivalent to maximize $\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t+1)}\pi_k + \sum_{(i,i')\in\mathscr{A}}\sum_{(k,k')=1}^{K} \tau_{ik}^{(t+1)}\tau_{i'k'}^{(t+1)} f(Y_{ii'}, \alpha_{kk'})$ i.e. the complete likelihood where the indicator functions on $\mathbf{Z}$ have been replaced with the parameters τ of the variational distribution. For classical distributions $\mathscr{F}$ on $\mathbf{Y}$, the update is explicit.

We presented the highlights of the VEM inference for the specific case of SBM. This algorithm can be adapted to infer LBM. As in the EM algorithm the initialization has a major importance and shall be carefully chosen in pratice. On a theoretical note, the consistency of the VEM estimates has been established for the SBM in [12] and the LBM in [24] while the behavior of $\mathbb{P}(\mathbf{Z}|\mathbf{Y};\hat{\theta})$ is studied in [114] for the same models.

Selection of the number of blocks. Many model selection criteria such as AIC or BIC consist in penalizing the likelihood. An Integrated Classification Likelihood (ICL) has been proposed in [13]. It has proven its capacity to outline the clustering structure in networks in [50] for simple networks, [94] for bipartite networks or [115] for valued networks. Its success comes from the fact that when traditional model selection criteria essentially involve a trade-off between goodness of fit and model complexity, ICL values not only goodness of fit but also clustering sharpness. We provide below its expression in the case of an undirected binary network:

$$\mathrm{ICL}(K) = \log \ell_c(\hat{\theta}_K; \mathbf{Y}, \hat{\mathbf{Z}}) - \mathrm{pen}(K) \tag{3.12}$$

where

$$\mathrm{pen}(K) = \frac{1}{2}\left\{(K-1)\log(n) + (K(K+1)/2)\log(n(n-1)/2)\right\}.$$

The term $\hat{\mathbf{Z}}$ may correspond to either the maximum a posteriori block recovery issued from the distribution $\mathscr{R}_{\mathbf{Y},\tau}(\mathbf{Z})$ or to their approximated posterior expectations given by the parameters τ defined in Equation (3.10). We then choose K such that ICL(K) is maximum.

3.4.2 Contributions

My contributions in the statistical analysis of networks are of two kinds. Some of them consist in extending the SBM to different kinds of multilayer networks [JP8, JP11, P6, P2]. These models were inferred on motivating datasets issued from sociology and ecology. They provided us with in-depth joint analyses of all the layers. The other one deals with the consideration of the network sampling effect when inferring an SBM [JP5].

Multilayer Networks There is a growing interest in multilayer networks. The term multilayer may refer to a large number of cases (see [136] for a review in ecology). It ranges from dynamic networks [95] that I have not considered in my work, to multiplex, multilevel and multipartite networks that I worked on. We define below multiplex, multilevel and multipartite networks and show how we extended the SBM to handle them. The identifiability of the extended SBMs was proven. For each extension, we adapted the VEM algorithm and the ICL criterion for inference and the selection of the number of blocks. Simulation studies were performed to assess the efficiency of the inference and the model selection. Furthermore, the extensions were motivated by datasets that we present below.

3.4.2.1 Multiplex Networks

In a multiplex network, several edges accounting for different kind of relationship between nodes occur. For instance, in a social network, two individuals may be tied

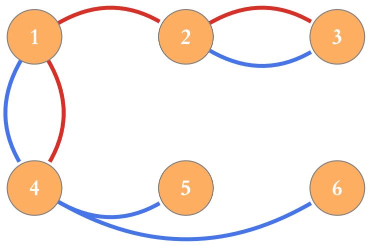


Figure 3.8 – *Illustration of a multiplex network. For each dyad, two kinds of link may exist. They are respectively displayed by red and blue edges.*

by a professional relationship, friendship or both (see Figure 3.8 for an illustration). One may expect that these two possible relationships are not independent. More interestingly, the dependence between the two relationships may vary from a group of individuals to another. Therefore, fitting an SBM with a multivariate Bernoulli distribution on dyads for multiplex networks will help to identify these different groups. More precisely, we assume that there are Q possible kinds of relationships between two nodes. Therefore, the variables $\mathbf{Y}_{ij} \in \{0,1\}^Q$ encode the effective edges in the network corresponding to the elements of the vectors being 1. In the multiplex SBM, Equation (3.6) is replaced with

$$\mathbf{Y}_{ij}|Z_i, Z_j \overset{ind}{\sim} \text{Bern}^Q((\alpha^w_{Z_i,Z_j})_w) \tag{3.13}$$

where Bern^Q is a multivariate Bernoulli distribution in dimension Q, $w \in \{0,1\}^Q$ and the parameters $\alpha^w_{Z_i,Z_j}$ correspond to the probabilities $\mathbb{P}(\mathbf{Y}_{ij} = w|Z_i, Z_j) = \alpha^w_{Z_i,Z_j}$. The parameters are such that: $\sum_w \alpha^w_{Z_i,Z_j} = 1$ resulting in $2^Q - 1$ free parameters to estimate per couple of blocks.

In [JP8], we posited this model, proved its identifiability and developed the dedicated VEM inference and number of blocks selection through an adapted ICL criterion. Note that we encoded the inference for the multiplex SBM for $Q = 2$ in the `R package blockmodels` [109].

Applications. The multiplex extension of the SBM was motivated in [JP8] by an application to an advice network between French cancer researchers. These data come from E. Lazega [107] who studied the relations of advice between French cancer researchers identified as “Elite” conjointly with the relations of their respective laboratories. Since the focus is on Elite researchers, it is almost a one-to-one correspondance between researchers and labs. That is why we can consider the lab relationship as an undirect relationship between researchers. Therefore, the researcher network is a multiplex network on which we fitted our multiplex SBM. Figure 3.9 displays the marginal and conditional (given the existence or not of an undirect relation through the labs) probabilities of a direct relation between researchers belonging to the 4 different blocks. It shows that the existence of a connection (exchange of resources) between labs clearly increases the probability of connection (sharing advice) between researchers. The reinforcement of this probability of connection is clearly outstanding in block 2. In this block, the researcher connections are quite unlikely within the block or with other blocks. However, conditionally to the existence of a laboratory connection, the researcher connections become more important especially with block 4. Researchers in block 3 seem to be the least affected by the connections provided by their laboratories.

This clustering demonstrates that not all researchers benefit on equal terms from the institutional level. Some researchers are more dependent on their laboratories in terms of connections.

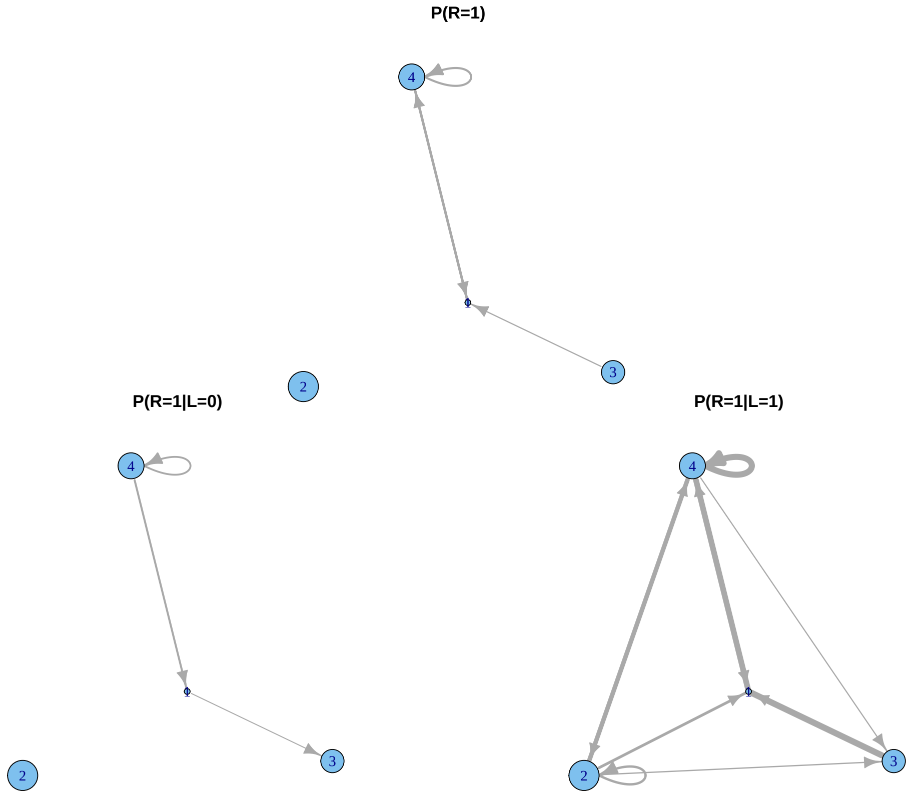


Figure 3.9 – *Marginal probabilities of Researcher connections between and within blocks (top) and probabilities of Researcher connections between and within blocks conditionally on absence (bottom left-hand-side) or presence (bottom right-hand-side) of Lab connection. Node size is proportional to the block size. Edge width is proportional to the probabilities of connection; if this probability is smaller than 0.1, edges are not displayed.*

In another collaboration with Emmanuel Lazega, we applied the multiplex SBM on a dataset which is still focused on the French cancer researchers but where the two possible relations are advice and competition [JP11]. In addition to identifying from whom the researchers seek advice, they were asked who they considered their competitors. Interestingly, we were able to highlight that most researchers take the risk of seeking advice from colleagues whom they identify as direct competitors.

3.4.2.2 Multilevel Networks

Multilevel networks arise in sociology of organizations and collective action when willing to study jointly the social network of individuals and the interaction network of organizations the individuals belong to. Indeed, the individuals not only interact with each others but are also members of interacting organizations. Following [108], one

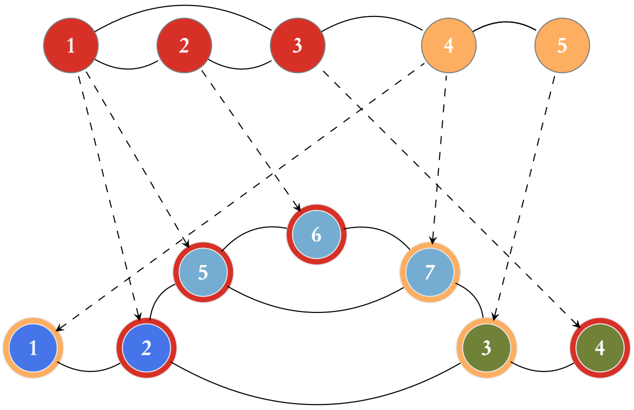


Figure 3.10 – *Illustration of a multilevel network following an MLVSBM. Inter-organizational level is on the top and inter-individual level is on the bottom. The various shades of blue depict the blocks of individuals and the various shades of red depict the blocks of organizations. The outer circles around the nodes of the individuals represent the blocks of the organizations they are affiliated to. The dashed links stand for the affiliations.*

might think that these two types of interactions (between individuals and between organizations) are interdependent, the individuals shaping their organizations and the organizations having an influence on the individuals. We aim to propose a statistical model for multilevel networks in order to understand how the two levels are intertwined and how one level impacts the other.

In [P2], we proposed the extension Multilevel SBM (MLVSBM) which can handle jointly the inter-individual and the inter-organisational networks and which relies on the affiliation matrix to make the two networks interdependent. We start by introducing the notations, then we provide the MLVSBM.

Let us consider n_I individuals involved in n_O organizations. We encode the networks into two adjacency matrices as follows. Let $\mathbf{Y}^I$ be the $n_I \times n_I$ matrix representing the inter-individual network and let $\mathbf{Y}^O$ be $n_O \times n_O$ matrix representing the inter-organizational network. These matrices may be symmetric or not, with loop or not, binary or valued. Moreover, the two matrices may be of different nature. For instance, $\mathbf{Y}^O$ may be a symmetric binary matrix and $\mathbf{Y}^I$ may be an asymmetric valued matrix. In addition to the two interaction matrices, we have the affiliation matrix A which is a $n_I \times n_O$ binary matrix such that:

$$A_{ij} = \begin{cases} 1 & \text{if individual i belongs to organization j,} \\ 0 & \text{otherwise} \end{cases}.$$

Moreover, A is contrained on its rows: $\forall i = 1, \ldots, n_I$, $\sum_{j=1}^{n_O} A_{ij} = 1$ since we assume that any individual belongs to a unique organization. A synthetic example of a multilevel network is given in Figure 3.10.

The MLVSBM is a blockmodel where individuals and organisations are clustered into blocks (respectively K_I and K_O blocks). The blocks of organisation are given by random variables $\mathbf{Z}^O$ such that: for all $j \in \{1, \ldots, n_O\}$, $k \in \{1, \ldots, K_O\}$,

$$\mathbb{P}(Z_j = k) \overset{ind}{=} \pi_k^O. \tag{3.14}$$

The dependence between the two levels is modeled through the assumption that the memberships of the individuals $\mathbf{Z}^I$ depend on the blocks of the organizations ($\mathbf{Z}^O$) they are affiliated to. More precisely, for all $i \in \{1,\ldots,n_I\}$, $k \in \{1,\ldots,K_I\}$,

$$\mathbb{P}(Z_i^I = k | Z_j^O, A_{ij} = 1) \overset{ind}{=} \gamma_{kZ_j^O}, \tag{3.15}$$

where γ is a $K_I \times K_O$ matrix such that $\sum_{k=1}^{K_I} \gamma_{kl} = 1 \; \forall l \in \{1,\ldots,K_O\}$. Conditionally on the latent variables $\mathbf{Z}^O$, $\mathbf{Y}^O$ follows a regular SBM with connection parameters in the matrix α^O and similarly for $\mathbf{Y}^I$ conditionally on $\mathbf{Z}^I$ with parameters in α^I.

The identifiability of the MLVSBM is proven under fairly general assumptions and we derive the following proposition stating in which cases the two levels are independent.

Proposition 4. *In the MLVSBM, the two following properties are equivalent:*

1. $\mathbf{Z}^I$ *is independent on* $\mathbf{Z}^O$*,*
2. $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1,\ldots,K_O\}$

and imply that:

3. $\mathbf{Y}^I$ *and* $\mathbf{Y}^O$ *are independent.*

Then, the VEM inference and the selection of the number of block can be derived.

Application. The dataset dealt with in the previous section on French cancer researchers corresponds originally to a multilevel network. However, the fact that most labs contain only one researcher makes the MLVSBM not really suited to these data. We infer the MLVSBM on another dataset concerned with the economic network of audiovisual firms and the informal network of their sales representatives during a television program trade fair [23]. These data were collected by face-to-face interviews. At the individual level, people were asked to select from a list the individuals from which they obtain advice or information during or before the trade fair. This level consists of 128 individuals who were affiliated to 109 organizations, each one containing from one to six individuals. At the inter-organizational level, the deal network (deals between organizations signed since the last trade fair) was collected.

In Figure 3.11, we provide synthetic views of the inferred MLVSBM. These representations are useful to understand how the networks are structured and to unravel some particular features. In [P2], the results are commented in depth. On a quick note, we point out that the loop in block 3 of individuals is an unexpected interesting structure that was unraveled thanks to the MLVSBM. Indeed, most of the relations happen between nodes from different blocks since the blocks correspond to groups of sellers or groups of buyers as confirmed by additional available covariates on the individuals. The loop in block 3 has an interesting sociological interpretation since it shows that in spite of competition between their firms, some individuals are still exchanging advice which was deemed as a *coopetition*.

3.4.2.3 Generalized Multipartite Networks

Multipartite networks are a generalization of bipartite networks. In a bipartite network, the nodes (representing the interacting entities) are partitioned into two disjoint

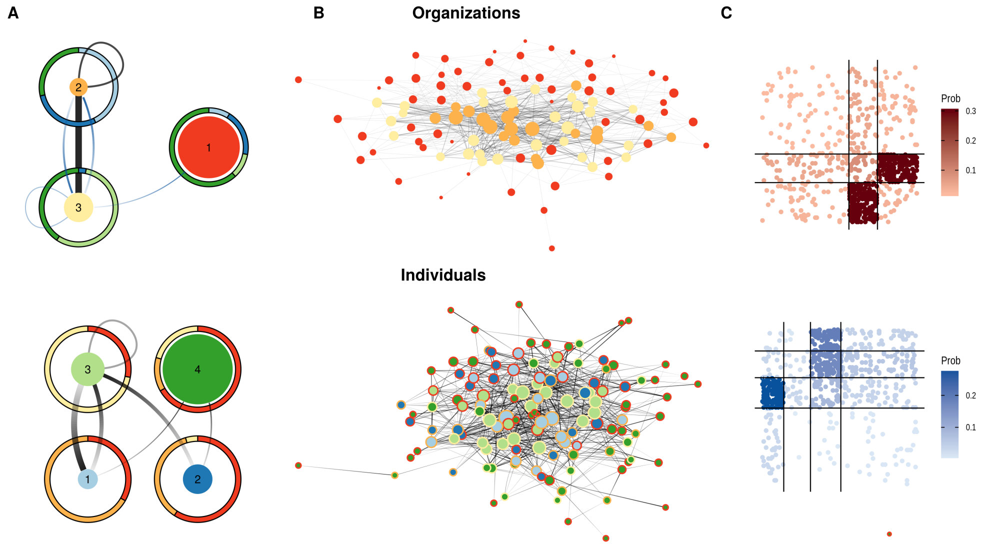


Figure 3.11 – *Multilevel network of the Promoshow East trade fair 2011. Top: the deal network for the organizations and bottom: the advice network for the individuals.* ***A:*** *Mesoscopic view of the multilevel network. Nodes stand for the blocks, donut charts show the relation between* $\mathbf{Z}^O$ *and* $\mathbf{Z}^I$*. Black edges are the probabilities of connection* α^I *and* α^O*, blue edges stand for* $\mathbb{P}(Y^I_{ii'} = 1|Z^O_{A_i}, Z^O_{A_{i'}})$*. For sake of clarity only edges with probabilities above the density are shown.* ***B:*** *View of the network. The size of a node is proportional to its centrality degree. Colors represent the clustering obtained with the multilevel SBM.* ***C:*** *Adjacency matrices of the advice network between individuals and the deal network between organizations. Entries are reordered by block from left to right and top to bottom.*

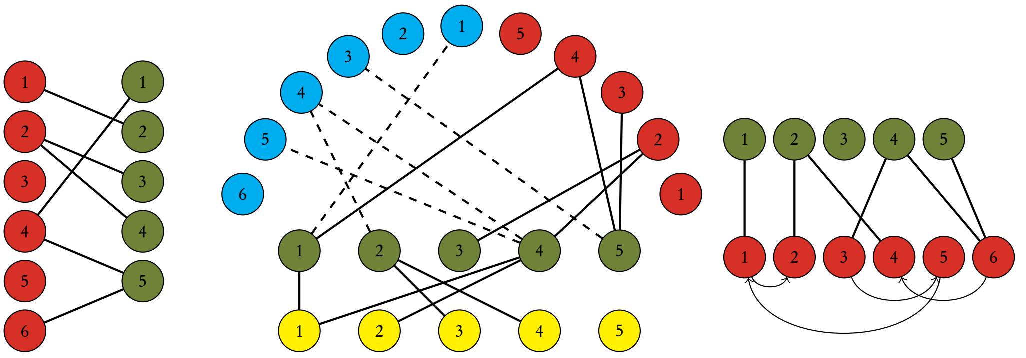


Figure 3.12 – *Illustrations of bipartite (left), multipartite (center) and generalized multipartite networks (right). The colors stand for the different functional groups.*

sets and an edge links a node from one set to a node from the other set (see Figure 3.12, left). In a multipartite network, the nodes are divided into more than two sets and edges link entities from different sets (see Figure 3.12, middle). In what follows, these pre-specified sets of nodes will be referred to as *functional groups*. Such multipartite networks arise in ecology when studying the interactions between several groups of species such as the interactions plant/pollinator, plants/ants, etc [138, 49] or in biology when analyzing networks issued from multi-omics datasets involving proteins, etc. [134]. Generalized multipartite networks are an extension of multipartite networks: the nodes are still partitioned into functional groups but the interactions may occur not only between different functional groups but also within some of the functional groups (see Figure 3.12, right).

More notations are needed to define the dataset corresponding to a multipartite network. We consider that there are Q functional groups and within each functional group, there are n_q individuals ($q = 1,\dots,Q$). The collection of networks is indexed by pairs of functional groups (q,q') (q and q' in $\{1,\dots,Q\}$). The set $\mathscr{L}$ denotes the list of pairs of functional groups for which we observe an interaction network. For any $(q,q') \in \mathscr{L}$, the interaction network is encoded in a matrix $Y^{qq'}$. The generalized multipartite network is then the collection of networks $\mathbf{Y} = \left(Y^{qq'}\right)_{(q,q')\in\mathscr{L}}$. For each network, $\mathscr{S}^{qq'}$ is an additional notation which refers to the list of all the possible interactions. This extension was motivated by two applications:

1. The **dataset 1** is issued from [49]. This ecological network gathers mutualistic relations between plants and pollinators, plants and ants, and plants and frugivorous birds, resulting into $Q = 4$ functional groups, namely plants ($q = 1$), pollinators ($q = 2$), ants ($q = 3$) and birds ($q = 4$) and $\mathscr{L} = \{(1,2),(1,3),(1,4)\}$. $Y^{1q'}_{ii'} = 1$ if the plant species i has been observed at least once in a mutualistic interaction with the animal species i' of functional group q' during the observation period, 0 otherwise.

2. The **dataset 2** comes from [163] and [164]. They collected the oriented network of seed circulation between farmers –resulting in a non-symmetric adjacency matrix – and the crop species grown by the farmers, resulting in an incidence matrix. Noting $q = 1$ for the farmers and $q = 2$ for crop species we get $\mathscr{L} = \{(1,1),(1,2)\}$. $Y^{11}_{ii'} = 1$ if farmer i gives seeds to farmer i' (oriented relation), 0 otherwise and $Y^{12}_{ij} = 1$ if farmer i cultivates crop species j, 0 otherwise.

To extend the SBM to the multipartite block model (MBM) which is able to handle generalized multipartite network, we assume a block clustering within each functional group. Then, for each interaction network $Y^{qq'}$ with $(q,q')\in\mathscr{L}$, we assume either an SBM if $q'=q$ or an LBM otherwise. More precisely, $\forall q=1,\ldots,Q$, $\forall i\in\{1,\ldots,n_q\}$, $\forall k\in\{1,\ldots,K_q\}$:

$$\mathbb{P}(Z_i^q=k)\overset{ind}{=}\pi_k^q.$$

Then, $\forall(i,i')\in\mathscr{S}^{qq'}$,

$$Y_{ii'}^{qq'}|\{Z_i^q,Z_{i'}^{q'}\}\overset{ind}{\sim}\mathscr{F}_{qq'}(\alpha^{qq'}_{Z_i^q Z_{i'}^{q'}}). \tag{3.16}$$

This model is a generalization of the SBM and the LBM. Indeed, the previous equations reduce to the SBM if $\mathscr{L}=\{(1,1)\}$ and to the LBM if $\mathscr{L}=\{(1,2)\}$. Our extension assumes that the latent structures $\mathbf{Z}$ are shared among the $Y^{qq'}$ i.e. if a functional group q is at stake in several $Y^{qq'}$, the same $\mathbf{Z}^q$ impacts the distributions of the corresponding interaction matrices. In other words, the clusters gather individuals sharing the same properties of connection in the full collection of networks. Obviously, if each functional group appears in only one element of $\mathscr{L}$, the MBM reduces to independent SBMs or LBMs. The distribution $\mathscr{F}_{qq'}$ are indexed by the functional groups in interaction since we may assume different distributions for the interaction matrices.

We derived a VEM algorithm and an ICL model selection criterion in [P6]. In the MBM, the practical choice of the number of blocks is computationally intensive since if we assume that $K_q\in\{1,\ldots,K_q^\star\}$ for all q, then we should compare $\prod_{q=1}^Q K_q^\star$ models through the ICL criterion. For each model, the VEM algorithm has to be run starting from a large number of initialization points chosen carefully (due to its sensitivity to the starting point), resulting in an unreasonable computational cost. Instead, we propose to adopt a stepwise strategy, resulting in a faster exploration of the model space combined with efficient initializations of the VEM algorithm. The procedure we suggest is given in Algorithm 3. It is implemented in the `R package GREMLIN` [R1].

Applications. The inference and the model selection were performed on the two datasets presented above. Figures 3.13 and 3.14 show respectively mesoscopic representations of datasets 1 and 2. By summarizing the multipartite networks, they offer a global vision of the interactions at stake. We compared the blocks issued by the MBM with blocks issued by SBM or LBM inferred on the different interaction matrices. We show that the joint inference provided by the MBM results in more complex blocks since the inference relies on more information. An alternative could be to infer separate SBM or LBM and then create the intersection of outputted blocks. However, this approach would result in too many blocks while our MBM provides a trade-off between the number of blocks and the available information.

3.4.3 SBM inference from Sampled Data

The inference and the model selection were presented in Section 3.4.1.2 under the assumption that the sampling of $\mathbf{Y}$ is complete. Here, we consider cases of binary networks where all the nodes are observed but information regarding the presence/absence of an edge is missing for some dyads. In other words the adjacency matrix contains missing values, a situation often met with real-world networks. For instance

Algorithm 3: Model selection strategy for MBM

```
Initialization Starting from a model $\mathcal{M}^{(0)} = \mathcal{M}(K_1^{(0)}, \dots, K_Q^{(0)})$.
Iterations
Given a current model $\mathcal{M}^{(m)} = \mathcal{M}(K_1^{(m)}, \dots, K_Q^{(m)})$, the $m$-th iteration is:
• Split proposals. For any $q$ such that $K_q^{(m)} < K_q^\star$, consider the model
      $\mathcal{M}_+^{(m+1)q} = \mathcal{M}(K_1^{(m)}, \dots, K_q^{(m)} + 1, \dots, K_Q^{(m)})$.
    · Propose $K_q^{(m)}$ initializations by splitting any of the $K_q^{(m)}$ current clusters into two clusters.
    · From each of the $K_q^{(m)}$ initialization points, run the VEM algorithm and keep the better variational estimate $\hat{\theta}_{\mathcal{M}_+^{(m+1)q}}$.
    · Compute the corresponding ICL($\mathcal{M}_+^{(m+1)q}$).
• Merge proposals. For any $q$ such that $K_q^{(m)} > 1$, consider the model
      $\mathcal{M}_-^{(m+1)q} = \mathcal{M}(K_1^{(m)}, \dots, K_q^{(m)} - 1, \dots, K_Q^{(m)})$.
    · Propose $K_q^{(m)}(K_q^{(m)} - 1)/2$ initializations by merging any pairs of clusters among the $K_q^{(m)}$ clusters.
    · From each initialization point, run the VEM algorithm and keep the better variational estimate $\hat{\theta}_{\mathcal{M}_-^{(m+1)q}}$.
    · Compute the corresponding ICL($\mathcal{M}_-^{(m+1)q}$).
• Set $\mathcal{M}^{(m+1)} = \arg\max_{\mathbb{M}^{(m)}} \mathrm{ICL}(\mathcal{M})$ where
$\mathbb{M}^{(m)} = \{\mathcal{M}^{(m)}\} \cup \bigcup_{q \in \{1, \dots Q\}} \{\mathcal{M}_+^{(m+1)q}\} \cup \{\mathcal{M}_-^{(m+1)q}\}$.
If $\mathcal{M}^{(m+1)} \neq \mathcal{M}^{(m)}$ iterate, otherwise stop.
```

in social sciences, network data consists in interactions between individuals: the set of individuals is fixed, possibly known from a census. Information about the presence/absence of an edge is only available when at least one of the two individuals is available for an interview, otherwise it is missing (see [96, 80] for a review of sampling techniques).

Sampled networks. From the adjacency matrix, we may derive for convenience an $n \times n$ binary matrix $\mathbf{R}$ such that $R_{ij} = 1$ if Y_{ij} is observed (i.e. we know if there is or not an edge on the dyad (i, j), respectively $Y_{ij} = 1$ or $Y_{ij} = 0$), $R_{ij} = 0$ otherwise

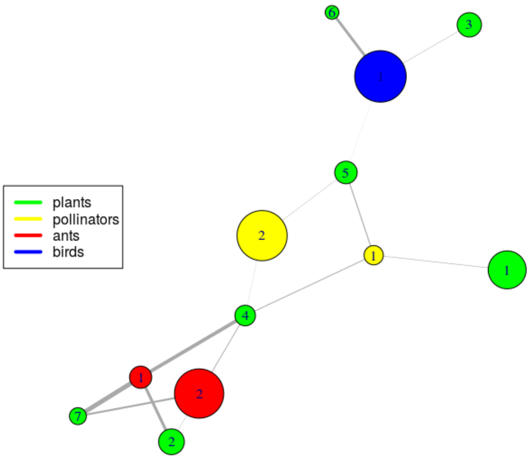


Figure 3.13 – *Mesoscopic view of dataset 1. Nodes stand for the inferred blocks, their size are proportional to the size of the blocks and the width of the edges are proportional to the probability of connection between/within blocks. Edges corresponding to probabilities of connection lower than 0.01 are not plotted.*

($Y_{ij} = \mathtt{NA}$). The natural idea is to replace Equation 3.8 with:

$$\log \ell_c(\theta; \mathbf{Y}, \mathbf{Z}) = \sum_{i \in \mathcal{N}^o} \sum_{k=1}^{K} \mathbb{1}_{\{Z_i = k\}} \pi_k + \sum_{(i,i') \in \mathcal{D}^o} \sum_{(k,k')=1}^{K} \mathbb{1}_{\{Z_i = k, Z_{i'} = k'\}} f(Y_{ii'}, \alpha_{kk'}) \quad (3.17)$$

where $\mathcal{N}^o = \{i \in \{1, \ldots, n\}, s.t. \sum_j R_{ij} + \sum_j R_{ji} > 0\} \subset \mathcal{N}$ the set of nodes involved in at least an observed dyad and $\mathcal{D}^o = \{(i,j) \in \mathcal{A} : R_{ij} = 1\}$ the set of observed dyads. Considering this complete likelihood is correct under the assumption that the sampled data are Missing At Random (MAR) defined in the seminal work of D. Rubin on missing data [144]. Indeed, the joint distribution of the SBM and the sampling mechanism are separable in the MAR case. Then, the SBM likelihood is optimized on the observed data only.

In [JP5], we also dealt with Not Missing At Random (NMAR) sampling mechanisms and derived the VEM inference. In these cases, the joint complete likelihood which incorporates the sampling and the SBM on the full network must be considered. Before giving the inference method, we present the three usual types of missingness (MCAR: Missing Completely At Random, MAR and NMAR) for SBM. This typology depends on the relations between the adjacency matrix $\mathbf{Y}$, the latent structure $\mathbf{Z}$ and the sampling $\mathbf{R}$, so that the missingness is characterized by four directed acyclic graphs displayed in Figure 3.15. MCAR samplings include random dyad and random node samplings. In random dyad sampling, each dyad has the same probability, say ρ to be observed independently of the others while in random node sampling, the nodes are sampled with the same probability independently of the others. Sampling a node means observing all the dyads in which it is involved, i.e. if node i is

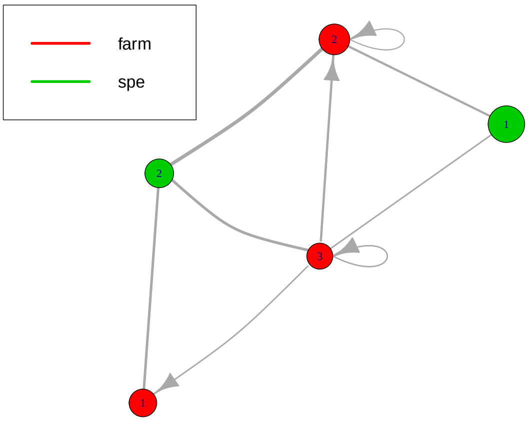
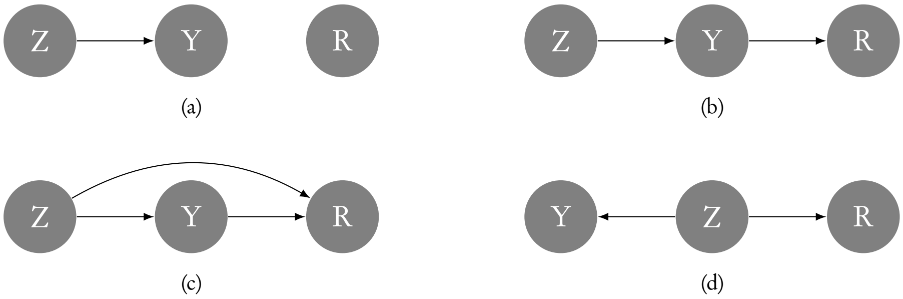


Figure 3.14 – *Mesoscopic view of dataset 2. Nodes stand for the inferred blocks, their size are proportional to the size of the blocks and the width of the edges are proportional to the probability of connection between/within blocks. farm stands for farmers and spe for species. The probability of connection below* 0.2 *are not plotted.*



(a)

(b)

(c)

(d)

Figure 3.15 – *DAGs of relationships between* $\mathbf{Y}, \mathbf{Z}$ *and* $\mathbf{R}$ *in the framework of missing data for SBM. DAG where* $\mathbf{R}$ *is a parent node are not reviewed since the network exists before the sampling design acts upon it. The systematic edge between* $\mathbf{Z}$ *and* $\mathbf{Y}$ *is due to the definition of the SBM. Note that the DAG* (b) *may correspond to MAR or NMAR samplings.*

sampled we have $R_{ij} = R_{ji} = 1$ for all j. The distinction between MAR and MCAR samplings is subtle. In MCAR, the sampling is totally independent on the network while in MAR, one may consider a sequential sampling such as snowball sampling. The snowball sampling starts with a first batch of nodes. All the dyads involving these nodes are observed. Then, the second step consists of observing the nodes (with all their dyads) connected to the first batch. These steps are called waves and are repeated a few times leaving potentially some nodes unobserved. In [JP5], we detailed some particular NMAR samplings which could be encountered in practical situations. We present them below.

Definition 1 (Double standard sampling)**.** *Let* $\rho_1, \rho_0 \in [0,1]$*. Double standard sampling consists in observing dyads with probabilities*

$$\mathbb{P}(R_{ij} = 1 | Y_{ij} = 1) = \rho_1, \qquad \mathbb{P}(R_{ij} = 1 | Y_{ij} = 0) = \rho_0. \tag{3.18}$$

Denote $S^o = \sum_{(i,j)\in\mathcal{D}^o} Y_{ij}$, $\bar{S}^o = \sum_{(i,j)\in\mathcal{D}^o}(1-Y_{ij})$ and similarly for $S^m, \bar{S}^m$ where the superscript m indicates the sum over the missing dyads. In this dyad-centered sampling design satisfying DAG (b), the log-likelihood is

$$\log \ell(\psi;\mathbf{R}|\mathbf{Y}) = S^{\mathrm{o}} \log \rho_1 + \bar{S}^{\mathrm{o}} \log \rho_0 + S^{\mathrm{m}} \log(1-\rho_1) + \bar{S}^{\mathrm{m}} \log(1-\rho_0), \quad \text{with } \psi = (\rho_0, \rho_1). \tag{3.19}$$

This sampling is likely to happen especially in cases with $\rho_1 > \rho_0$ which means that it is easier to detect an existing link than its absence.

Definition 2 (Star sampling based on degrees – Star degree sampling). *Star degree sampling consists in observing all dyads corresponding to nodes selected with probabilities* $\{\rho_1, \ldots, \rho_n\}$ *such that* $\rho_i = \text{logistic}(a + bD_i)$ *for all* $i \in \mathcal{N}$ *where* $(a,b) \in \mathbb{R}^2$, $D_i = \sum_j Y_{ij}$ *and* $\text{logistic}(x) = (1+e^{-x})^{-1}$.

In this node-centered sampling design satisfying DAG (b), the log-likelihood is

$$\log \ell(\psi;\mathbf{R}|\mathbf{Y}) = \sum_{i \in \mathcal{N}^{\mathrm{o}}} \log \rho_i + \sum_{i \in \mathcal{N}^{\mathrm{m}}} \log(1-\rho_i), \qquad \text{with } \psi = (a,b). \tag{3.20}$$

In this sampling, we assume that the degree is somehow related to the popularity of a node and that this popularity makes them more likely to be sampled (when $b > 0$).

Definition 3 (Class sampling). *Class sampling consists in observing all dyads corresponding to nodes selected with probabilities* $\{\rho_1, \ldots, \rho_Q\}$ *such that* $\rho_q = \mathbb{P}(i \in \mathcal{N}^{\mathrm{o}} \mid Z_i = k)$ *for all* $(i,q) \in \mathcal{N} \times K$.

In this node-centered sampling design satisfying DAG (d), the log-likelihood is

$$\log \ell(\psi;\mathbf{R}|\mathbf{Z}) = \sum_{i \in \mathcal{N}^{\mathrm{o}}} \sum_{q \in \mathcal{Q}} Z_{iq} \log \rho_q + \sum_{i \in \mathcal{N}^{\mathrm{m}}} \sum_{q \in \mathcal{Q}} Z_{iq} \log(1-\rho_q), \quad \text{with } \psi = (\rho_1, \ldots, \rho_Q). \tag{3.21}$$

Here, we assume the blocks stand for different communities and that the sampling does not have the same ability to reach them.

We proved under mild conditions the identifiability of two MCAR samplings (dyad and node centered samplings) and of the class sampling. The identifiability concerns at the same time the sampling parameters and the SBM parameters. Moreover, the consistency and asymptotic normality of VEM estimators are proven in [116] for the MCAR dyad centered sampling.

Inference for NMAR situations. The VEM algorithm must be adapted for NMAR cases by incorporating the sampling likelihood. Therefore, the lower bound of the likelihood which we aim to maximize is no longer the one given in Equation 3.11 but it is:

$$\mathcal{J}_\theta(\mathcal{R}_{(\mathbf{Y}^{\mathrm{m}},\mathbf{Z})}) = \log \ell(\theta, \psi; \mathbf{Y}^{\mathrm{o}}, \mathbf{R}) - \mathrm{KL}\left[\mathcal{R}_{(\mathbf{Y}^{\mathrm{m}},\mathbf{Z})}, \mathbb{P}(\cdot|\mathbf{Y}^{\mathrm{o}}; \theta, \psi)\right].$$

where ψ stands for the sampling parameters. The variational distribution $\mathcal{R}_{(\mathbf{Y}^{\mathrm{m}},\mathbf{Z})}$ concerns both the missing dyads $\mathbf{Y}^{\mathrm{m}}$ and the latent variable $\mathbf{Z}$. We seek for $\mathcal{R}_{(\mathbf{Y}^{\mathrm{m}},\mathbf{Z})}$ in the class of independent distribution:

$$\mathcal{R}_{(\mathbf{Y}^{\mathrm{m}},\mathbf{Z})} = \mathcal{R}_{(\mathbf{Y}^{\mathrm{m}})} \cdot \mathcal{R}_{(\mathbf{Z})} = \prod_{(i,j) \in \mathcal{D}^{\mathrm{m}}} \nu_{ij}^{Y_{ij}} (1-\nu_{ij})^{1-Y_{ij}} \cdot \prod_{i \in \mathcal{N}} \prod_{k=1}^{K} (\tau_{ik})^{\mathbb{1}_{Z_i=k}}, \tag{3.22}$$

where the ν_{ij}s and τ_{ik}s are the parameters to be optimized in the VE step of the algorithm. The optimization in ν_{ij} is specific to the sampling design and must be derived for each case while the optimization in τ_{ik} is almost generic. Similarly, in the M step the optimization is generic in θ and specific to the sampling in ψ. Concretely, when optimizing in τ_{ik} or in θ the update formulas are similar to the full observed situation. The NA in $\mathbf{Y}^{\mathrm{m}}$ are replaced with their variational corresponding parameters ν_{ij}. In [JP5], the complete formulas were derived for the three NMAR samplings presented above. An ICL criterion may be derived not only to select the number of blocks K (as for fully observed network, see Equation (3.12)) but also for selecting the most appropriate sampling design when it is unknown. In particular, this criterion may be used to decide whether an MAR sampling fits better the data than an NMAR sampling. For an undirected network, the ICL criterion is

$$\mathrm{ICL}(K) = \log \ell_c(\hat{\theta}_K, \hat{\psi}; \mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{R}, \hat{\mathbf{Z}}) - \mathrm{pen}(K) \tag{3.23}$$

where

$$2\mathrm{pen}(K) = \begin{cases} \left(d + \frac{K(K+1)}{2}\right)\log\left(\frac{n(n-1)}{2}\right) + (K-1)\log(n) & \text{for dyad-centered sampling} \\ \frac{K(K+1)}{2}\log\left(\frac{n(n-1)}{2}\right) + (d+K-1)\log(n) & \text{for node-centered sampling} \end{cases},$$

where d is the dimension of ψ the vector of sampling parameters.

In Figure 3.16, we show on synthetic data the benefits of taking into account the NMAR sampling in the inference over MAR inference when the NMAR sampling is actually responsible for the missing data. The adapted NMAR VEM inference has almost a perfect recovery of blocks and estimation of α while the errors in MAR VEM inference increase as the difference $\rho_1 - \rho_0$ moves away from 0.

Importance of accounting for missing values in real networks. We dealt with two datasets where there are missing data on the adjacency matrix. We compared the clusterings obtained under the assumptions of an MAR or an NMAR samplings. It has been shown that there are big differences and the nature of the sampling cannot be overlooked. The first dataset is concerned with seed exchanges of sorghum in the region of Mount Kenya. The data were collected and analyzed in [102, 101]. The sampling is node-centered since the exchanges are documented by interviewing farmers who are asked to declare to whom they gave seeds and from whom they receive seeds. Since an interview is time consuming, the sampling is not exhaustive. A limited space area was defined where all the farmers were interviewed. The network is thus collected with missing dyads since information on the potential links between two farmers who were cited but not interviewed is missing. Since we only know that the sampling is node-centered, we fitted SBM under the three node-centered sampling designs random node sampling (MAR), class and star degree samplings (NMAR). The ICL criterion is minimal for 10 blocks under the star degree sampling and for 11 blocks under the class degree sampling. The clusterings between the SBMs obtained with either class or star degree sampling remain close from each other and both unravel a strong community structure. The adjusted Rand Index (ARI) [139] between these two clusterings is 0.6. The model selected by ICL for MAR sampling is composed by 11 blocks. The ARIs between MAR clustering and the two others are lower (around 0.4). Although, the ICL criteria computed for the three sampling designs are a slightly in favor of the MAR sampling, the clusterings under NMAR conditions are more connected to the social structure. The social structure is indeed given by additional covariates on nodes which

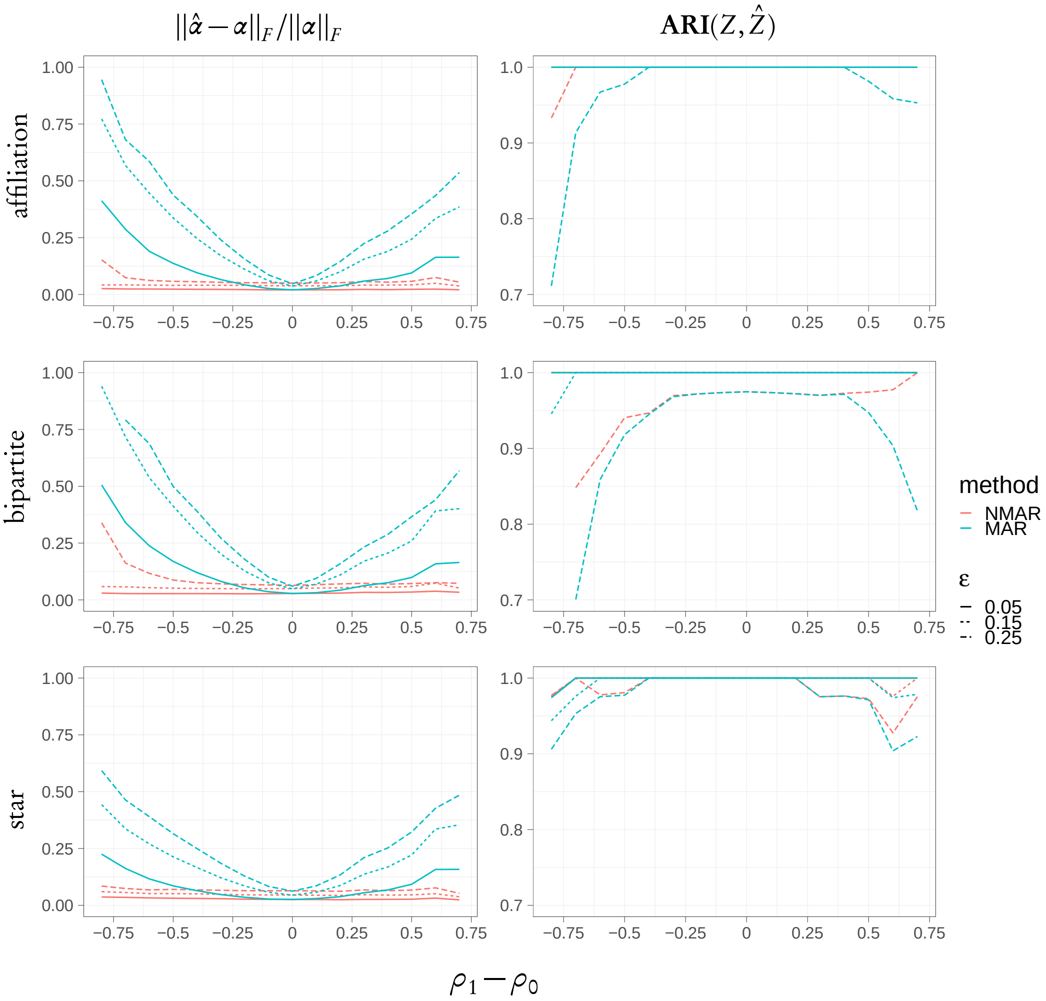


Figure 3.16 – *Double standard setting: estimation error of α and adjusted Rand index averaged over 500 simulations for affiliation, bipartite and star topologies.*

are the dialect spoken by the farmer and the neighborhood where they live. We highlighted that these two covariates have a stronger relation with the clusterings issued under NMAR samplings than with the one under an MAR sampling.

The second dataset we consider is a Protein-Protein Interaction network which encodes possible relations between proteins. We extracted the interaction network of the neighborhood of the Estrogen receptor protein from the platform `string` [161]. It provides a valued network. The values are scores $\omega_{ij} \in [0,1]$ which indicate how likely is an interaction between a pair of proteins. We cast this network in the framework of binary network with missing data on dyads by choosing a threshold γ such that the network is:

$$\mathbf{Y}^{\gamma} = (Y^{\gamma})_{ij} = \begin{cases} 1 & \text{if } \omega_{ij} > 1-\gamma, \\ \texttt{NA} & \text{if } \gamma \leq \omega_{ij} \leq 1-\gamma, \\ 0 & \text{if } \omega_{ij} < \gamma. \end{cases} \tag{3.24}$$

We fixed the threshold $\gamma = 0.35$ and fitted an SBM under random dyad sampling (MAR) and double standard sampling (NMAR). The two corresponding SBMs have 11 clusters for MAR sampling and 13 clusters for NMAR sampling. The ARI between the two clusterings is around 0.39: this is mainly due to a large block in the random-dyad MAR clustering which contains much more nodes than any of the blocks in the NMAR clustering. The latter dispatches many of these nodes in four blocks. We relied on the Gene Ontology annotation [4] to prove that this finest clustering of the nodes is more relevant from the biological point of view.

Taking into account covariates. In T. Tabouy's Ph. D. thesis [162], we also considered the extension when covariates are available and may impact both the distribution of the random network **Y** and the sampling process **R**. Conditional dependencies of the possible models are represented with a DAG in Figure 3.17. We choose to study only cases where covariates represented by **X** impact either the sampling design or the network model directly. Finally, we do not consider any edge between nodes **Y**, **Z** and node **R** since we require that conditionally on **X** the missing data are MAR. Some models were proposed for each DAG in [162] and we showed an interesting equivalence between a case which is MAR provided that the covariates are taken into account and its NMAR counterpart when the covariates are not observed. The details of the inference algorithms for the proposed models are also in the manuscript.

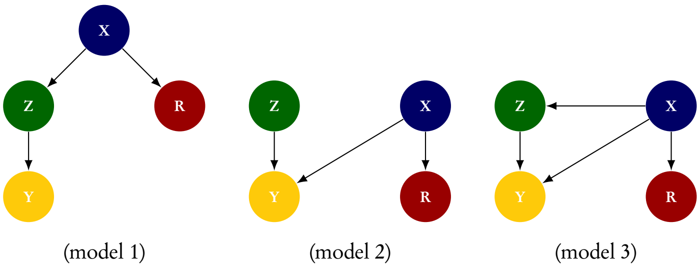


Figure 3.17 – *DAGs of relationships between* **Y**, **Z**, **R** *and* **X** *considered in the framework of missing data for SBM with covariates.*

All the inference algorithms and the corresponding ICL criterion computations for the SBM with the different samplings described above (with or without covariates) are implemented in the `R package missSBM` [R2]. A software paper describing the package and providing some examples of use is available [P5].

3.5 Perspectives

My perspectives on networks can be grouped in three main themes. The first one is concerned with the inference of the underlying network which impacts a diffusion process as in Sections 3.2 and 3.3. Since the problem is complex, the inference may be limited to coarse descriptions of the network. The second perspective is about the sampling effect when inferring a random graph model on a network and also an efficient inference of the missing data relying on all available information. Finally, the third perspective is focused on the derivation and the analysis of specific summary statistics such as robustness in ecology, under blockmodel assumption for the interaction network.

3.5.1 Inferring Networks in a Dynamic Model

Inferring network features from genetic and demographic data. During the internship of Sixtine de Cussac (AgroParisTech, spring 2016), we explored the possibility to recover the social mechanism which structures the seed circulation from data on genetic diversity in the field. We assumed a dynamic genetic metapopulation model in discrete time for n farms (nodes) which grow plants during m time steps. A generation (transition from a time step to the next one) for each farm consists in the steps plotted in Figure 3.18. First an extinction takes place with a fixed probability. If the extinction does not occur, the next generation of plants in this farm is simulated from a genetic reproduction model. Otherwise, the seeds are recovered either from a neighboring farm (model M1) or from the market (model M2) where all farmers put their seeds in common. The neighbors are defined on the basis of an observed contact network among farmers. We assumed a very simple genetic reproduction model where we focused on a unique biallelic locus. The goal was to decide whether model M1 or M2 is at play in the seed circulation from the observation of the neighboring network and the genetic data for the plants in all the farms at generation m. Since the likelihood is intractable, standard statistical decision methods were not possible. We resorted to simulation-based methods in the spirit of Approximate Bayesian computation methods [117]. We used the simulation model under M1 and M2 to create a large dataset on which we learnt a statistical decision rule to decide which of M1 or M2 is the most likely with respect to the genetic data. Then, we could apply this decision rule on real data. Obviously, the genetic data are too large. They need to be summarized into small dimension statistics. It makes sense to use diversity indices such as the fixation index (FST) [174] or diversity indices such as β-diversity [171]. As it was expected, these indices allowed us to separate M1 from M2 since M2 has a homogenizing effect by redistributing seeds from all the farms to the farms which go extinct. Since the data are too scarce, it is unrealistic to try to infer the whole network. We rather focused on coarse information such as simply determining whether the network was effectively used to recover seed. Estimating additional parameters such as the extinction rate and the parameters driving the reproduction would make the decision harder.

The question of assessing to what extent the social relationships among farmers shape the cultivated diversity is an emerging issue. The work done during the internship was a very first step in this direction. We could consider more complex genetic data with markers both under selection or not and less assumptions on the neighboring network. Thus, the goal could be to infer some topological features of the networks from the data. To do so, we aim to use CropMetaPop which is an agent-based model build on the Python simuPOP [135]. It combines the evolution of plants represented

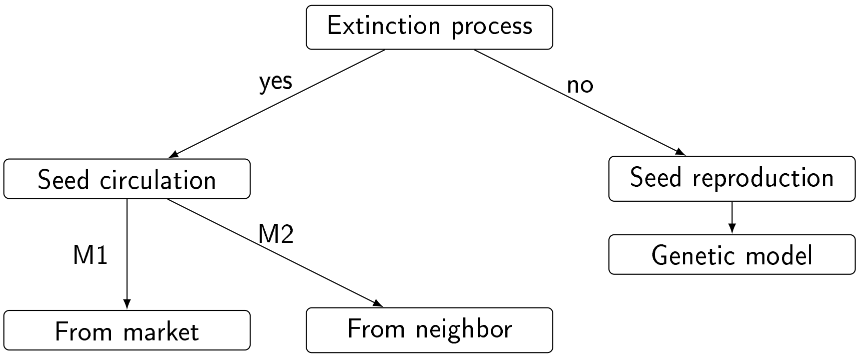


Figure 3.18 – *One generation of the dynamic model for a farm with seed circulation according to models M1 or M2.*

as multilocus genomes in several populations which interact according to a social network.

These works make some interesting connection with the perspectives on stochastic simulators (see Section 2.3.2). Indeed, the metapopulation genetic models are stochastic simulators. Inferring their inputs on the basis of some observations can be cast to a calibration task. We may resort to emulation in order to make the calibration possible as an alternative to ABC methods.

Inferring networks. The inference of the contact network is reachable provided that the full dynamic for all nodes is observed as described in Section 3.3. The approach could be extended to more complex propagation models, the propagation rules being encoded in the terms ϕ_{ij}^t as in Equation (3.4). More than two levels of infection ('sick' or 'healthy') can be considered: when the contamination duration is known, the model can be extended to an SIR model, by simply adding a 'recovered' state. In the same vein, the effect of environmental factors could also be accounted for via a regression term in the transition rates encoded in the ϕ_{ij}^t. The difficulty of the parameter inference will mostly depend on the expression of ϕ_{ij}^t, but the complexity of the network reconstruction will remain the same and will still benefit from the computational efficiency achieved through the Matrix-tree theorem. Additional information on contacts could also be encoded in the parameters β_{ij} (Equation (3.3)) as prior information on possible contacts. For instance, if some information on distances between individuals is available, a parametric form for β_{ij} linking the probability of contact to the distance could be assumed.

The network could also be available at a coarser scale, the nodes representing countries or geographical area and not individuals. The data would be the number of infected cases, recovered cases and the goal would be to identify the main paths of contamination.

3.5.2 Sampled Networks

Inference for other NMAR samplings. Natural perspectives from the contribution of Section 3.4.3 are to develop inference for other NMAR samplings, some of them related to distribution for **Y** different from the Bernoulli distribution. In particular, extending this contribution to DCSBM or PABM is sensible since it may be assumed that the probability to observe a node depends on its degree, popularity. We encountered many cases in our simulation studies where the inference dedicated to an MAR case was still performing well in spite of the NMAR sampling. On a theoretical note, we aim to investigate to what extent the inference under MAR condition remains robust.

Sampling effect in ecological interaction networks. In ecology, sampling a network is labor-intensive and many datasets only reveal a subset of the existing interactions. Obviously, the sampling process can induce enormous biases in the statistical analyses of the networks which have not been taken into account in most papers concerned with the analysis of ecological network structure. This raises doubts regarding the current understanding of the structure of pollination networks [19]. In particular, recent studies [159] suggest that observed nestedness in interaction networks mainly results from sampling effects. Moreover,the impact of the completeness of the sampling on many metrics such as modularity or nestedness is demonstrated in [142]. Completeness of sampling for a species is defined as the proportion of observed interactions in which that species is involved. This completeness may be evaluated either through accumulation curves which model the rate of new interaction observations over time [142] or through external data assessing the abundance of the different interacting species [18]. The natural idea for metrics to cope with sampling effect is to evaluate its significance with respect to null models that take into account this sampling effect. Such null models are obtained through resamplings of the networks which keep the marginal distribution of abundance (degrees of rows and columns) fixed. The null model is not unique and its choice may lead to different conclusions (see [58] for a software and a discussion on null models). Contrary to metrics, the blockmodels being probabilistic model, it is quite easy to incorporate the sampling effect within the model.

To illustrate the incorporation of the sampling effect, we use data provided in [18]. This dataset contains a plant-ant network which is obtained from counts of ant colonies attending extrafloral and floral nectaries of plants. Then, this is a weighted incidence matrix where rows correspond to ants species and columns correspond to extrafloral or floral nectaries of plant species. In addition to the incidence matrix, independent abundance estimates for ant colonies based on a sugar bait experiment and number of plant individuals on which any insect were recorded on nectaries are available as abundance estimate. This abundance may represent a sampling effect since the more abundant a species, the more complete its sampling. Indeed, an abundant species is more likely to have most of its interaction observed than a less abundant one. Therefore, a higher degree may correspond to two indistinguable situations if there is no external data on abundance: either the species is more abundant than the others or the species shares more connection than the others. On this data we fit three blockmodels. Model 1 corresponds to a classical LBM with a distribution on dyads as a Poisson distribution: $Y_{ij}|Z_i^1, Z_j^2 \overset{ind}{\sim} \mathcal{P}(\alpha_{Z_i^1,Z_j^2})$. Model 2 is a practical implementation of a degree correction in the LBM where a covariate is associated with each species: $Y_{ij}|Z_i^1, Z_j^2 \overset{ind}{\sim} \mathcal{P}(\alpha_{Z_i^1,Z_j^2}\mu_i\nu_j)$ where μ_i, $1 \leq i \leq n$ and $\nu_i, 1 \leq j \leq n$ are positive parameters and $\mu_1 = \nu_1 = 1$ for the sake of identifiability. The covariates account for a potential effect of the respective abundance of the species on the interaction. Then the parameters α should account for the network structure beyond the abundance effect. Eventually, model 3 incorporates the external abundance estimates as covariates: $Y_{ij}|Z_i^1, Z_j^2 \overset{ind}{\sim} \mathcal{P}(\exp(\alpha_{Z_i^1,Z_j^2} + \beta R_i + \delta S_j))$ where R_i is the independent abundance

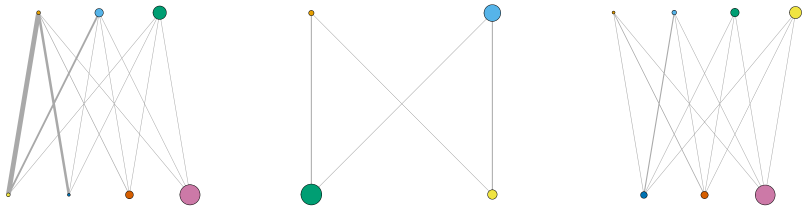


Figure 3.19 – *Summaries of the incidence network provided by the* 3 *models (from left to right: model* 1, 2 *and* 3*). Nodes in the upper level represent ants and plants in the lower level. Node sizes are proportional to block sizes. Edge widths are proportional to* λ *parameters of the Poisson distribution between blocks.*

measure for row i, S_j is the independent abundance measure for row j and β, δ are parameters to be estimated. Note that only the external abundance estimates have a medium correlation with row (ants) degrees (0.50) and a high correlation with column (plant) degrees (0.99). Therefore this external information is not totally redondant with the degree. The inferred LBM matrix are plotted in Figure 3.19, the numbers of blocks correspond to optimal choices according to ICL criterion. With model 1, we find a nested structure which is mainly driven by the relative abundance of species. The nested structure disappears when inferring model 2. The only structure that remains consists of two blocks of ants and two blocks of plants which are paired with each other (ants in the orange block are more likely to connect with plants in the green block and ants in the blue block with plants in the yellow block). The last model finds also something quite different. It consists in a mixture between a nested structure and some particular associations. The ARIs between the different clustering on ants or on plants outputed by the three inferred models were computed, leading to moderate values between 0.17 and 0.64. Eventually, taking into account the relative abundance makes a difference in the discovered underlying structure and this question should be carefully addressed when collecting or analyzing such data.

When such external abundance data are not available, the full observation data including the full observation times of each interaction may help to fit accumulation curves [129]. For instance, if the network at stake is a plant-pollinator network, we fit for each plant a model such as the Clench model: $S(t) = at/(1 + bt)$ where t is the time of observation and S the number of observation of pollinators (from different species). The parameters a and b are estimated from the full data and their ratio gives the maximal number of pollinator for a given plant a/b. Then for each plant, we compute the completeness as the ratio between the observed number of pollinators and this theoretical maximal number of pollinators. These completeness scores may be used as above as covariates in an LBM.

Another solution could be to propose a model for the sampling process as done in Section 3.4.3. The major difference lies in the fact that the data $\mathbf{Y}$ for a sampled binary interaction network consists only of 1 and `NA`. There is no observed 0 since the absence of observation does not mean for sure that the interaction is not possible. This process is inherently NMAR which then requires dedicated inference algorithms. The sampling process may depend on some external abundance measure or on the completeness scores.

An exciting application for these works could be to analyze data from citizen science program where the sampling is not done by scientific campaign but by opportunistic sampling. For example, Spipoll (`www.spipoll.org`) aims to monitor plant-pollinator interactions in France. They gather more than 300,000 records across seasons by hikers or nature enthusiasts. These data are massive but their drawback is the sampling distribution. Some area are more popular than others, it is more likely to have people collecting data when the weather is nice and so on. Analyzing these data in spite of their specific sampling is a major challenge and is of obvious interest.

Missing dyad reconstruction through different layers. When observing a multilayer network (multiplex, multipartite, multilevel), the quality of data collection may differ between layers. For instance, interaction data for a given layer may be easier to collect or may be already extensively documented. Then, under an MAR sampling assumption, this well observed layer may help in many ways. In multipartite networks, since the latent blocks are assumed to be shared among the layers, the identification of latent blocks is made more robust. In multiplex networks, besides a better identification of the latent blocks, the interdependencies between the layers help to predict the dyads in the missing layers. In [52], the authors propose to use one or more layers for predicting some missing dyads in another layer. They also assess the interdependencies between layers by computing the improvement in the missing dyad prediction due to the other layers. However, although the multilayer networks they consider are actually multiplex networks, they do not model the dependence between dyads from different layers beyond the one induced by the common latent blocks. Using our MBM (Section 3.4.2.3) we could extend this approach to generalized multipartite network and using our multiplex SBM (Section 3.4.2.1) for multiplex networks, we could improve the missing dyad prediction and have a more accurate interdependency assessment through our more complex modeling of dependence between dyads. In [P2], we used this idea to assess the interdependence between the two levels.

The other layers through our multilayer networks not only help to improve the prediction of missing dyads but also to correct spurious information as suggested in [38] or [79]. Spurious information in a network is a result of an error in the data collection. The dyads are not labeled as missing but incorrectly observed. In such situations, the other layers could help to cast doubt on some observations which seem unlikely and then prompt to check.

3.5.3 Robustness

A key question in ecological networks is their ability to withstand perturbations [59, 6], for instance, the effect on the whole network of the extinction of a species. A typical analysis in ecological networks consists in removing a species (node) in a network and recording the subsequent additional (referred to as ‘secondary’) extinctions. This number of secondary extinctions is then considered as a stability metric of the network, called robustness. In plant-pollinator networks, the primary extinctions may concern the pollinator. Then, the secondary extinctions are counted as the plants which are no longer connected to any pollinator. The pollinator species are removed sequentially and the cumulative number of secondary extinctions among plants are plotted. A robustness index is then computed from this curve. It may be either the area under curve or the number of extinctions among pollinator which leads to halve the number of plants. The robustness index depends on the sequence of removed pollinator species. Therefore, the robustness indices based on many sequences are averaged to remove this dependence.

An alternative idea is to fit a blockmodel on the interaction network. From the inferred blockmodel, the mean robustness is tractable analytically. Here, we consider a binary interaction network $\mathbf{Y} \in \{0,1\}^{p\times n}$ with p plants divided into K_1 blocks, n pollinators divided into K_2 blocks. We assume that $\mathbf{Y}$ follows an LBM with parameters $(\pi^1_q)_{1\leq q\leq K_1}$ for the marginal block distribution of plants, $(\pi^2_q)_{1\leq q\leq K_2}$ for the marginal block distribution of pollinators and connection parameters $\alpha \in [0,1]^{K_1\times K_2}$. If m pollinators are removed uniformly, for each plant i we compute the probability that it goes extinct as

$$\begin{aligned}
\mathbb{P}(\cap_{j=1}^{n-m}(Y_{ij}=0)) &= \sum_{q=1}^{K_1}\mathbb{P}(\cap_{j=1}^{n-m}(Y_{ij}=0|Z_i^1=q)\cdot\mathbb{P}(Z_i^1=q),\\
&= \sum_{q=1}^{K_1}\left(\prod_{j=1}^{n-m}\mathbb{P}(Y_{ij}=0|Z_i^1=q)\right)\cdot\mathbb{P}(Z_i^1=q),\\
&= \sum_{q=1}^{K_1}\left(1-\sum_{l=1}^{K_2}\pi_l^2\alpha_{ql}\right)^{n-m}\pi_q^1.
\end{aligned}$$

Therefore, the mean number of secondary extinction among plants for m extinctions of pollinators is

$$p\cdot\sum_{q=1}^{K_1}\left(1-\sum_{l=1}^{K_2}\pi_l^2\alpha_{ql}\right)^{n-m}\pi_q^1.$$

By summing up over m, we obtain the area under the curve of the cumulative secondary extinctions. This results incorporates both the variability of the sequence of pollinator extinction and the variability over realizations of the LBM with these parameters. These computations will help to compare different structure of LBM with respect to their associated robustness in which ecologists have a major interest. Moreover, computing the robustness from the LBM may be an interesting way not to rely too much on the sampled network which remains an imperfect observation of the structuring interactions. The structure of the network captured by the LBM lies mainly in the parameters π^1, π^2 and the parameter α up to a constant. Indeed, the number of plants, pollinators and even the global density are contingent to the particular observation and hence to the sampling. We can derive robustness with different values for n, p and the global density. By doing so, we can then cope with some sampling effects and produce fairer comparisons between networks. For instance, if we assume that the sampling may miss nearly one half of possible interactions but we still assume that we are able to infer the structure of the LBM (π^1, π^2 and the parameter α up to a constant), we can leverage the LBM assumption by doubling α and computing the robustness with this adjusted parameter. The comparison of robustness between networks of different sizes (number of pollinators) or of different densities (which may result of a sampling effort) is not always fair since these two features have a major impact. Under the LBM assumption, we can adjust parameters in order to cast the networks in the same setting where they only differ on their structures leading therefore to a fair comparison.

Finally, all these works can be extended to the cases of multilayer networks with the same kind of exact computation under different assumptions of blockmodels. In this framework, the cascading extinction will be the major focus [26].

BIBLIOGRAPHY

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25, 2000.

[5] F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013.

[6] M. S. Bane, M. J. Pocock, and R. James. Effects of model choice, network structure, and interaction strengths on knockout extinction models of ecological robustness. *Ecology and evolution*, 8(22):10794–10804, 2018.

[7] M. Bayarri, J. Berger, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R. Parthasarathy, R. Paulo, J. Sacks, and D. Walsh. Computer model validation with functional output. *The Annals of Statistics*, pages 1874–1906, 2007.

[8] M. Bayarri, J. Berger, R. Paulo, J. Sacks, J. Cafeo, J. Cavendish, C. Lin, and J. Tu. A framework for validation of computer models. *Technometrics*, 49(2):138–154, 2007.

[9] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, 2012.

[10] J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[11] K. S. Bhat, D. S. Mebane, P. Mahapatra, and C. B. Storlie. Upscaling uncertainty with dynamic discrepancy for a multi-scale carbon capture system. *Journal of the American Statistical Association*, 112(520):1453–1467, 2017.

[12] P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943, 08 2013.

[13] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, Jul 2000.

[14] M. Binois, R. B. Gramacy, and M. Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 0(0):1–14, 2018.

[15] M. Binois, R. B. Gramacy, and M. Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 0(0):1–14, 2018.

[16] M. Binois, J. Huang, R. B. Gramacy, and M. Ludkovski. Replication or exploration? sequential design for stochastic simulation experiments. *Technometrics*, 0(0):1–17, 2018.

[17] P. Block. Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40:163–173, 2015.

[18] N. Blüthgen, N. E Stork, and K. Fiedler. Bottom-up control and co-occurrence in complex communities: Honeydew and nectar determine a rainforest ant mosaic. *Oikos*, 106(2):344–358, 2004.

[19] N. Blüthgen, J. Fründ, D. P. Vázquez, and F. Menzel. What do interaction network metrics tell us about specialization and biological traits. *Ecology*, 89(12):3387–3399, 2008.

[20] C. Bouveyron, P. Latouche, and R. Zreik. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 28(1):11–31, 2018.

[21] G. E. Box and N. R. Draper. *Empirical model-building and response surfaces.* John Wiley & Sons, 1987.

[22] G. E. P. Box and G. T. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, 1973.

[23] J. Brailly, G. Favre, J. Chatellet, and E. Lazega. Embeddedness as a multilevel problem: A case study in economic sociology. *Social Networks*, 44:319–333, 2016.

[24] V. Brault, C. Keribin, and M. Mariadassou. Consistency and Asymptotic Normality of Latent Blocks Model Estimators. preprint, Apr. 2017.

[25] J. Brynjarsdóttir and A. O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse problems*, 30(11):114007, 2014.

[26] S. Buldyrev, R. Parshani, G. Paul, H. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025–1028, 2010.

[27] C. Cannamela, J. Garnier, B. Iooss, et al. Controlled stratification for quantile estimation. *The Annals of Applied Statistics*, 2(4):1554–1580, 2008.

[28] G. Casella and E. Moreno. Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167, 2006.

[29] G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics: An International Journal of Probability and Stochastic Processes*, 41(1-2):119–134, 1992.

[30] A. Celisse, J.-J. Daudin, L. Pierre, et al. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

[31] S. Chaiken. A combinatorial proof of the all minors matrix tree theorem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):319–329, 1982.

[32] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security*, 10(4):1:1–1:26, 2008.

[33] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[34] A. Channarond. *Clustering in a random graph : models with latent space*. Theses, Université Paris Sud - Paris XI, 2013.

[35] M. Chelle and B. Andrieu. The nested radiosity model for the distribution of light within plant canopies. *Ecological Modelling*, 111(1):75–91, 1998.

[36] M.-H. Chen and Q.-M. Shao. On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, 25(4):1563–1594, 1997.

[37] C. Chevalier and D. Ginsbourger. Fast computation of the multi-points expected improvement with applications in batch selection. In *International Conference on Learning and Intelligent Optimization*, pages 59–69. Springer, 2013.

[38] A. Clauset, C. Moore, and M. E. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98, 2008.

[39] G. Consonni, E. Gutiérrez-Peña, and P. Veronese. Compatible priors for Bayesian model comparison with an application to the hardy–weinberg equilibrium model. *Test*, 17(3):585–605, 2008.

[40] S. Conti, J. P. Gosling, J. E. Oakley, and A. O'Hagan. Gaussian process emulation of dynamic computer codes. *Biometrika*, 96(3):663–676, 2009.

[41] N. R. Council et al. *Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification*. National Academies Press, 2012.

[42] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[43] P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith. Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In *Case studies in Bayesian statistics*, pages 37–93. Springer, 1997.

[44] N. Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

[45] S. Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

[46] G. Damblin. *Contributions statistiques au calage et à la validation des codes de calcul*. PhD thesis, Université Paris Saclay, Nov. 2015.

[47] G. Damblin, M. Couplet, and B. Iooss. Numerical studies of space-filling designs: optimization of latin hypercube samples and subprojection properties. *Journal of Simulation*, 7(4):276–289, 2013.

[48] A. Darwinkel. Patterns of tillering and grain production of winter wheat at a wide range of plant densities. *Netherlands Journal of Agricultural Science*, 1978.

[49] W. Dáttilo, N. Lara-Rodríguez, P. Jordano, P. R. Guimarães, J. N. Thompson, R. J. Marquis, L. P. Medeiros, R. Ortiz-Pulido, M. A. Marcos-García, and V. Rico-Gray. Unravelling Darwin's entangled bank: architecture and robustness of mutualistic networks with multiple interaction types. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1843), 2016.

[50] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008.

[51] J. R. Day and H. P. Possingham. A stochastic metapopulation model with variability in patch size and position. *Theoretical Population Biology*, 48:333–360, 1995.

[52] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317, 2017.

[53] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

[54] B. Delyon, M. Lavielle, E. Moulines, et al. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[55] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[56] S. Donnet and A. Samson. Estimation of parameters in incomplete data models defined by dynamical systems. *J. Statist. Plann. Inference*, 137:2815–2831, 2007.

[57] S. Donnet and A. Samson. Parametric inference for mixed models defined by stochastic differential equations. *ESAIM: Probability and Statistics*, 12:196–218, 2008.

[58] C. F. Dormann, J. Fründ, N. Blüthgen, and B. Gruber. Indices, graphs and null models: analyzing bipartite ecological networks. *The open ecology journal*, 2009.

[59] J. A. Dunne, R. J. Williams, and N. D. Martinez. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology letters*, 5(4):558–567, 2002.

[60] S. Duretz, J.-L. Drouet, P. Durand, N. J. Hutchings, M. Theobald, J. Salmon-Monviola, U. Dragosits, O. Maury, M. Sutton, and P. Cellier. Nitroscape: a model to integrate nitrogen transfers and transformations in rural landscapes. *Environmental Pollution*, 159(11):3162–3170, 2011.

[61] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[62] K.-T. Fang, R. Li, and A. Sudjianto. *Design and modeling for computer experiments*. Chapman and Hall/CRC, 2005.

[63] A. Flachs, G. D. Stone, and C. Shaffer. Mapping knowledge: GIS as a tool for spatial modeling of patterns of warangal cotton seed popularity and farmer decision-making. *Human ecology*, 45(2):143–159, 2017.

[64] M. A. Fortuna, D. B. Stouffer, J. M. Olesen, P. Jordano, D. Mouillot, B. R. Krasnov, R. Poulin, and J. Bascompte. Nestedness versus modularity in ecological networks: two sides of the same coin? *Journal of animal ecology*, 79(4):811–817, 2010.

[65] M. Frenklach, A. Packard, G. Garcia-Donato, R. Paulo, and J. Sacks. Comparison of statistical and deterministic frameworks of uncertainty quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):875–901, 2016.

[66] S. Fu, G. Celeux, N. Bousquet, and M. Couplet. Bayesian inference for inverse problems occurring in uncertainty analysis. *International Journal for Uncertainty Quantification*, 5(1), 2015.

[67] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[68] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[69] L. J. Gilarranz and J. Bascompte. Spatial network structure and metapopulation persistence. *Journal of Theoretical Biology*, 297(0):11 – 16, 2012.

[70] N. Gilbert. *Agent-based models*. Number 153. Sage, 2008.

[71] L. Gilquin, T. Capelle, E. Arnaud, and C. Prieur. Sensitivity analysis and optimisation of a land use and transport integrated model. *Journal de la Société Française de Statistique*, 2016.

[72] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[73] P. W. Goldberg, C. K. Williams, and C. M. Bishop. Regression with input-dependent noise: A Gaussian process treatment. In *Advances in neural information processing systems*, pages 493–499, 1998.

[74] G. Govaert and M. Nadif. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, 39(3):416–425, 2010.

[75] R. B. Gramacy and H. K. Lee. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012.

[76] M. Gu and L. Wang. Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1555–1583, 2018.

[77] M. Gu, X. Wang, J. O. Berger, et al. Robust Gaussian stochastic process emulation. *The Annals of Statistics*, 46(6A):3038–3066, 2018.

[78] J. Guedj, R. Thiébaut, and D. Commenges. Maximum likelihood estimation in dynamical models of HIV. *Biometrics*, 63:1198–2006, 2007.

[79] R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

[80] M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, 2010.

[81] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[82] I. Hanski and O. Ovaskainen. The metapopulation capacity of a fragmented landscape. *Nature*, 404:755–758, 2000.

[83] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[84] R. Herbei and L. M. Berliner. Estimating ocean circulation: An MCMC approach with approximated likelihoods via the bernoulli factory. *Journal of the American Statistical Association*, 109(507):944–954, 2014.

[85] D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

[86] D. Higdon, M. Kennedy, J. Cavendish, J. Cafeo, and R. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

[87] B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering & System Safety*, 94(7):1194–1204, 2009.

[88] M. E. Johnson, L. M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of statistical planning and inference*, 26(2):131–148, 1990.

[89] D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[90] K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau. Testing hypotheses via a mixture estimation model. *arXiv preprint arXiv:1412.2044*, 2014.

[91] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[92] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[93] M. Kennedy and A. O'Hagan. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, 63(3):425–464, 2001.

[94] C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015.

[95] B. Kim, K. H. Lee, L. Xue, X. Niu, et al. A review of dynamic network models with latent variables. *Statistics Surveys*, 12:105–135, 2018.

[96] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[97] B. A. Konomi, G. Karagiannis, K. Lai, and G. Lin. Bayesian treed calibration: an application to carbon capture with AX sorbent. *Journal of the American Statistical Association*, 112(517):37–53, 2017.

[98] D. G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

[99] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[100] K. N. Kyzyurova, J. O. Berger, and R. L. Wolpert. Coupling computer models through linking their statistical emulators. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1151–1171, 2018.

[101] V. Labeyrie, M. Deu, A. Barnaud, C. Calatayud, M. Buiron, P. Wambugu, S. Manel, J.-C. Glaszmann, and C. Leclerc. Influence of ethnolinguistic diversity on the sorghum genetic patterns in subsistence farming systems in eastern kenya. *PLoS One*, 9(3):e92178, 2014.

[102] V. Labeyrie, M. Thomas, Z. K. Muthamia, and C. Leclerc. Seed exchange networks, ethnicity, and sorghum diversity. *P. Natl. Acad. Sci.*, 113(1):98–103, 2016.

[103] M. Lamboni, D. Makowski, S. Lehuger, B. Gabrielle, and H. Monod. Multivariate global sensitivity analysis for dynamic crop models. *Field Crops Research*, 113(3):312–320, 2009.

[104] D. B. Larremore, A. Clauset, and A. Z. Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

[105] P. Latouche, E. Birmelé, C. Ambroise, et al. Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.

[106] M. Lavielle, A. Samson, A. Fermin, and F. Mentre. Maximum likelihood estimation of long term HIV dynamic models and antiviral response. *Biometrics*, 67(1):250–259, 2011.

[107] E. Lazega, M.-T. Jourda, L. Mounier, and R. Stofer. Catching up with big fish in the big pond? multi-level network analysis through linked design. *Social Networks*, 30(2):159 – 176, 2008.

[108] E. Lazega and T. A. Snijders. *Multilevel network analysis for the social sciences: Theory, methods and applications*, volume 12. Springer, 2015.

[109] J.-B. Léger. blockmodels. `https://cran.r-project.org/web/packages/blockmodels/index.html`, 2015.

[110] R. Levins. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the ESA*, pages 237–240, 1969.

[111] C. Linkletter, D. Bingham, N. Hengartner, D. Higdon, and K. Q. Ye. Variable selection for Gaussian process models in computer experiments. *Technometrics*, 48(4):478–490, 2006.

[112] F. Liu, M. Bayarri, and J. Berger. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.

[113] F. Liu, M. West, et al. A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Analysis*, 4(2):393–411, 2009.

[114] M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 02 2015.

[115] M. Mariadassou, S. Robin, C. Vacher, et al. Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, 4(2):715–742, 2010.

[116] M. Mariadassou and T. Tabouy. Consistency and asymptotic normality of stochastic block models estimators from sampled data. *arXiv preprint arXiv:1903.12488*, 2019.

[117] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[118] S. Marque-Pucheu, G. Perrin, and J. Garnier. An efficient dimension reduction for the Gaussian process emulation of two nested codes with functional outputs. *Computational Statistics*, pages 1–41, 2019.

[119] S. Marque-Pucheu, G. Perrin, and J. Garnier. Efficient sequential experimental design for surrogate modeling of nested codes. *ESAIM: Probability and Statistics*, 23:245–270, 2019.

[120] G. Matheron. *Traité de géostatistique appliquée. 1 (1962)*, volume 1. Editions Technip, 1962.

[121] C. Matias, T. Rebafka, and F. Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680, 2018.

[122] C. Matias and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, 47:55–74, 2014.

[123] J. Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer, 1975.

[124] M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[125] M. D. Morris and T. J. Mitchell. Exploratory designs for computational experiments. *Journal of statistical planning and inference*, 43(3):381–402, 1995.

[126] R. E. Morrison, T. A. Oliver, and R. D. Moser. Representing model inadequacy: A stochastic operator approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):457–496, 2018.

[127] L. Muchnik, S. Pei, L. C. Parra, S. D. Reis, J. S. Andrade Jr, S. Havlin, and H. A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3(1):1–8, 2013.

[128] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[129] A. Nielsen and J. Bascompte. Ecological networks, nestedness and sampling effort. *Journal of Ecology*, 95(5):1134–1141, 2007.

[130] M. Noroozi, R. Rimal, and M. Pensky. Sparse popularity adjusted stochastic block model. *arXiv preprint arXiv:1910.01931*, 2019.

[131] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[132] T. A. Oliver, G. Terejanu, C. S. Simmons, and R. D. Moser. Validating predictions of unobserved quantities. *Computer Methods in Applied Mechanics and Engineering*, 283:1310–1335, 2015.

[133] A. M. Overstall and D. C. Woods. Bayesian design of experiments using approximate coordinate exchange. *Technometrics*, 59(4):458–470, 2017.

[134] G. A. Pavlopoulos, P. I. Kontou, A. Pavlopoulou, C. Bouyioukos, E. Markou, and P. G. Bagos. Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience*, 7(4), 02 2018. giy014.

[135] B. Peng and M. Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, Sept. 2005.

[136] S. Pilosof, M. A. Porter, M. Pascual, and S. Kéfi. The multilayer nature of ecological networks. *Nature Ecology & Evolution*, 1(4):0101, 2017.

[137] M. Plumlee. Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285, 2017.

[138] M. J. Pocock, D. M. Evans, and J. Memmott. The robustness and restoration of a network of ecological networks. *Science*, 335(6071):973–977, 2012.

[139] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[140] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. the MIT press, 2006.

[141] B. Ribba, G. Kaloshi, M. Peyre, D. Ricard, V. Calvez, M. Tod, B. Cajavec-Bernard, A. Idbaih, D. Psimaras, L. Dainese, J. Pallud, S. Cartalat-Carel, J. Delattre, J. Honnorat, E. Grenier, and F. Ducray. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clin Cancer Res*, 18:5071–5080, 2012.

[142] A. Rivera-Hutinel, R. Bustamante, V. Marín, and R. Medel. Effects of sampling completeness on the structure of plant–pollinator networks. *Ecology*, 93(7):1593–1603, 2012.

[143] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[144] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[145] J. Sacks, W. W.J., T. Mitchell, and H. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[146] A. Saltelli, K. Chan, and E. Scott. *Sensitivity Analysis*. Wiley, New York, 2000.

[147] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity analysis*, volume 1. Wiley New York, 2000.

[148] A. Samson, M. Lavielle, and F. Mentré. The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Stat. Med.*, 26(27):4860–4875, 2007.

[149] T. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, 2003.

[150] T. Savitsky, M. Vannucci, and N. Sha. Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.*, 26(1):130–149, 2011.

[151] R. Schaback. Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264, 1995.

[152] R. Schaback. Kernel-based meshless methods. *Lecture Notes for Taught Course in Approximation Theory. Georg-August-Universität Göttingen*, 2007.

[153] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[154] S. Sengupta and Y. Chen. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):365–386, 2018.

[155] M. F. Sheridan, A. J. Stinton, A. Patra, E. Pitman, A. Bauer, and C. Nichita. Evaluating titan2d mass-flow model using the 1963 little tahoma peak avalanches, mount rainier, washington. *Journal of volcanology and geothermal research*, 139(1-2):89–102, 2005.

[156] T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[157] I. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp*, 1(4):407–414, 1993.

[158] R. V. Solé and J. Bascompte. *Self-Organization in Complex Ecosystems.* Princeton University Press, 2006.

[159] P. Staniczenko and et al. The ghost of nestedness in ecological networks. *Nature Com.*, 4:1391, 2013.

[160] E. R. Stefanescu, A. K. Patra, M. Bursik, E. B. Pitman, P. Webley, and M. D. Jones. Forecasting volcanic plume hazards with fast uq. *Procedia Computer Science*, 51:1613–1622, 2015.

[161] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic. Acids Res.*, 43, 2015.

[162] T. Tabouy. *Impact de l'échantillonnage sur l'inférence de structures dans les réseaux: application aux réseaux d'échanges de graines et à l'écologie.* PhD thesis, Paris Saclay, 2019.

[163] M. Thomas and S. Caillon. Effects of farmer social status and plant biocultural value on seed circulation networks in Vanuatu. *Ecology and Society*, 21(2), 2016.

[164] M. Thomas, N. Verzelen, P. Barbillon, O. T. Coomes, S. Caillon, D. McKey, M. Elias, E. Garine, C. Raimond, E. Dounias, et al. A network-based method to detect patterns of local crop biodiversity: validation at the species and infra-species levels. In *Advances in Ecological Research*, volume 53, pages 259–320. Elsevier, 2015.

[165] R. Tuo and C. Jeff Wu. A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795, 2016.

[166] R. Tuo and C. J. Wu. Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352, 2015.

[167] W. Ulrich, M. Almeida-Neto, and N. J. Gotelli. A consumer's guide to nestedness analysis. *Oikos*, 118(1):3–17, 2009.

[168] W. Ulrich and N. J. Gotelli. A null model algorithm for presence–absence matrices based on proportional resampling. *Ecological Modelling*, 244:20–27, 2012.

[169] B. Voight. A relation to describe rate-dependent material failure. *Science*, 243(4888):200–203, 1989.

[170] H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[171] R. H. Whittaker. Vegetation of the siskiyou mountains, Oregon and California. *Ecological monographs*, 30(3):279–338, 1960.

[172] D. Williamson, A. T. Blaker, C. Hampton, and J. Salter. Identifying and removing structural biases in climate models with history matching. *Climate dynamics*, 45(5-6):1299–1324, 2015.

[173] R. K. Wong, C. B. Storlie, and T. C. Lee. A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):635–648, 2017.

[174] S. Wright. Genetical structure of populations. *Nature*, 166(4215):247–249, 1950.

[175] H. Wu, Y. Huang, E. Acosta, S. Rosenkranz, D. Kuritzkes, J. Eron, A. Perelson, and J. Gerber. Modeling long-term HIV dynamics and antiretroviral response: effects of drug potency, pharmacokinetics, adherence, and drug resistance. *Journal of Acquired Immune Deficiency Syndromes*, 39:272–283, 2005.

[176] Q. Zhou, P. Z. Qian, and S. Zhou. A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53(3):266–273, 2011.