

## MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

---

La littérature empirique et thématique, ainsi que les cas d'études développés précédemment, semblent converger vers un consensus sur la complexité des relations entre réseaux de transport et territoires. Dans certaines configurations et à certaines échelles, il est possible de mettre en valeur des relations circulaires causales entre dynamiques territoriales et dynamiques des réseaux de transports. Nous désignons leur existence par le concept de *co-évolution*. Il semble difficile d'introduire des explications simples ou systématiques de ces dynamiques, comme le rappelle par exemple les débats autour des effets structurants des infrastructures [OFFNER, 1993].

Par ailleurs, les multiples situations géographiques suggèrent une forte dépendance au contexte, donnant une pertinence au travail de terrain et aux études ciblées. Or l'explication géographique et la compréhension des processus est très vite limitée dans cette approche, et intervient un besoin d'un certain niveau de généralisation. C'est sur un tel point que la théorie évolutive des villes se concentre en particulier, puisqu'elle permet de combiner des schémas et modèles généraux aux particularités géographiques. Au contraire, certaines théories issues de la physique s'appliquant à l'étude des systèmes urbains [WEST, 2017] peuvent être plus difficiles à accepter pour les géographes de par leur positionnement d'universalité qui est à l'opposé de leurs épistémologies habituelles.

Dans tous les cas, le *medium* qui permet de gagner en généralité sur les processus et structures des systèmes est toujours le modèle. Comme le rappelle J.P. MARCHAND<sup>1</sup>, "*notre génération a compris qu'il y avait une co-évolution, la votre cherche à la comprendre*", ce qui appuie le pouvoir de compréhension apporté par la modélisation et la simulation que nous jugeons être encore aujourd'hui à très fort potentiel de développement.

Sans développer pour le moment les nombreuses fonctions que peut avoir un modèle, nous nous baserons sur la position de BANOS qui soutient que "modéliser c'est apprendre", et suivant notre positionnement dans une science des systèmes complexes suggéré en introduction, nous ferons ainsi de la *modélisation des interactions entre réseaux et territoires* notre principal sujet d'étude, outil, objet<sup>2</sup>. Ce chapitre doit être pris comme un "état de l'art" des démarches

<sup>1</sup> Communication personnelle, Mai 2017.

<sup>2</sup> Même si dans une relecture à la lumière de 8.3 de ce positionnement n'a pas de sens puisque notre démarche contenait déjà des modèles à partir du moment où elle était scientifique.

de modélisation des interactions entre réseaux et territoires. Il vise en particulier à capturer différentes dimensions des connaissances : pour cela, nous mobiliserons des analyses en épistémologie quantitative.

Dans une première section 2.1, nous passons en revue de manière interdisciplinaire les modèles pouvant être concernés, même de loin, sans a priori d'échelle temporelle ou spatiale, d'ontologies, de structure, ou de contexte d'application. Cet aperçu est possible par les entrées disciplinaires diverses révélées au chapitre précédent : par exemple géographie, géographie des transports, planification. Cet aperçu suggère des structures de connaissances assez indépendantes et des disciplines ne communiquant que rarement.

Nous procédons dans 2.2 à une revue systématique algorithmique, qui correspond à une reconstruction par exploration itérative d'un paysage scientifique. Ses résultats tendent à confirmer ce cloisonnement. L'étude est complétée par une analyse de réseau multi-couches, combinant réseau de citations et réseau sémantique issu d'analyse textuelle, qui permet de mieux cerner les relations entre disciplines, leur champs lexicaux et leur motifs d'interdisciplinarité.

Cette étude permet la constitution d'un corpus utilisé pour la modélographie (typologie de modèles) et la méta-analyse (caractérisation de cette typologie) effectuée en dernière section 2.3. Celle-ci dis-  
sèque la nature d'un certain nombre de modèles et la relie au contexte disciplinaire, ce qui pose les bases et le cadre précis des efforts de modélisation qui seront développés par la suite.



*Ce chapitre est inédit pour sa première section ; reprend dans sa deuxième section le texte traduit de [RAIMBAULT, 2017d], puis pour sa deuxième partie la méthodologie introduite par [RAIMBAULT, 2016c] et développée dans [RAIMBAULT, 2017] ainsi que les outils de [BERGEAUD, POTIRON et RAIMBAULT, 2017a] ; et enfin est inédit pour sa dernière partie.*

## 2.1 MODÉLISER LES INTERACTIONS

### 2.1.1 *Modélisation en Géographie Quantitative*

#### *Histoire*

La modélisation joue en géographie théorique et quantitative un rôle fondamental. [CUYALA, 2014] procède à une analyse spatio-temporelle du mouvement de la géographie théorique et quantitative en langue française et souligne l'émergence de la discipline comme une combinaison d'analyses quantitatives (e.g. analyse spatiale et pratiques de modélisation et de simulation) et de construction théoriques. Cette dynamique est datée à la fin des années 1970, et est intimement liée à l'utilisation et l'appropriation des outils mathématiques [PUMAIN et ROBIC, 2002]. L'intégration de ces deux composantes permet la construction de théories à partir de faits stylisés empiriques, qui produisent à leur tour des hypothèses théoriques pouvant être testées sur les données empiriques. Cette approche est née sous l'influence de la *New Geography* dans les pays Anglo-saxons et en Suède.

Concernant la modélisation urbaine en elle-même, d'autres champs que la géographie ont proposé des modèles de simulation à peu près à la même période. Par exemple, le modèle de LOWRY, développé par [LOWRY, 1964] dans un but appliqué immédiat à la région métropolitaine de Pittsburg, suppose un système d'équations pour la localisation des actifs et des emplois dans différentes zones. Ce modèle a été une pierre angulaire de la modélisation urbaine, puisque comme le montre [GOLDNER, 1971] il avait déjà moins d'une dizaine d'années plus tard un conséquent héritage de développements conceptuels et opérationnels<sup>3</sup>. Des modèles relativement similaires sont toujours largement utilisés aujourd'hui.

#### *Simulation de modèle et calcul intensif*

Une histoire étendue de la genèse des modèles de simulation en géographie est faite par [REY-COYREHOURCQ, 2015] avec une attention particulière pour la notion de validation de modèles (nous reviendrons sur la place de ces aspects dans notre travail en 3). L'utilisation de ressources de calcul pour la simulation de modèles est antérieure à l'introduction des paradigmes de la complexité actuels, remontant par exemple à FORRESTER, informaticien qui a été pionnier des modèles d'économie spatiale inspirés par la cybernétique<sup>4</sup>. Avec l'augmentation des potentialités de calcul, des transformations

3 [GOLDNER, 1971] fait l'hypothèse que ce succès est du à la combinaison de trois facteurs : une possibilité d'application opérationnelle directe, une structure causale du modèle simple à appréhender (les actifs se localisent en fonction des emplois), et un cadre flexible pouvant être étendu ou adapté.

4 Celle-ci, ainsi que le courant systémique, sont comme nous l'avons déjà développé précurseurs des paradigmes actuels de la complexité.

épistémologiques ont également suivi, avec l'apparition de modèles explicatifs comme outils expérimentaux. REY compare le dynamisme des années soixante-dix quand les centres de calcul furent ouverts aux géographes à la démocratisation actuelle du Calcul Haute Performance<sup>5</sup>. Aujourd'hui, cette facilité d'accès consiste entre autres à du calcul sur grille dont l'utilisation est rendue transparente, c'est-à-dire sans besoin de compétences techniques pointues liées au mécanisme de la distribution des calculs. Ainsi [SCHMITT et al., 2015] donnent un exemple des possibilités offertes en termes de calibration et de validation de modèle, réduisant le temps de calcul nécessaire de 30 ans à une semaine - ces techniques jouent un rôle clé pour les résultats que nous obtiendrons par la suite. Cette évolution est également accompagnée par une évolution des pratiques [BANOS, 2013] et techniques [CHÉREL, COTTINEAU et REUILLON, 2015] de modélisation.

La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [LIVET et al., 2010] rappelle la combinaison des domaines empirique, conceptuel (théorique) et de la modélisation, avec des rétroactions constructives entre chaque. Un modèle peut être un outil d'exploration pour tester des hypothèses, un outil empirique pour valider une théorie sur des jeux de données, un outil explicatif pour révéler des causalités et ainsi des processus internes au système, un outil constructif pour construire itérativement une théorie conjointement avec celle des modèles associés. Ce sont des exemples de fonctions parmi d'autres : [VARENNE, 2010b] propose une classification des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein d'un système, plus particulièrement dans notre cas au sein d'un système complexe adaptatif. Nous rappelons ainsi que notre question de recherche s'intéressera aux *modèles dont l'ontologie contient une part non négligeable d'interactions entre réseaux et territoires*.

### 2.1.2 Modéliser les territoires et réseaux

Développons à présent un aperçu des différentes approches modélisant des interactions entre réseaux de transport et territoires. Remarquons de manière préliminaire une forte contingence des constructions scientifiques sous-jacentes à celles-ci. En effet, selon [BRETAGNOLLE, PAULUS et PUMAIN, 2002], *“les idées des spécialistes de la planification cherchant à donner des définitions des systèmes de ville, depuis 1830, sont étroitement liées aux transformations des réseaux de communication”*. Le contexte historique (et donc socio-économique et techno-

<sup>5</sup> Le développement des premiers modèles de simulation urbaine coïncide avec l'ouverture des premiers centres de calcul aux sciences humaines, comme le rappelle par ailleurs PUMAIN (entretien du 31/03/2017, voir Annexe D.3) par exemple pour l'implémentation du modèle d'entropie d'ALLEN.

logique) conditionne fortement les théories formulées. Cela implique que les ontologies et les modèles correspondants proposés par les géographes et les planificateurs sont fortement liés aux préoccupations historiques courantes, ce qui limite nécessairement leur portée théorique et/ou opérationnelle. Au delà de la question de la définition du système qui joue également un rôle central, on comprend bien l'impact que peut avoir cette influence sur la portée des modèles développés. Dans une vision perspectiviste de la science [GIERE, 2010c] de telles limites sont l'essence de l'entreprise scientifique, et comme nous suggérerons dans le chapitre 8 leur combinaison et couplage dans le cas de modèles est généralement une source de connaissance.

L'entrée que nous proposons ici pour dresser un aperçu des modèles est complémentaire à celle prise au chapitre 1, en regardant par objet principal (c'est-à-dire les relations Réseau  $\rightarrow$  Territoire, Territoire  $\rightarrow$  Réseau et Territoire  $\leftrightarrow$  Réseau)<sup>6</sup>.

Le cadre de lecture des échelles est également celui proposé au chapitre 1, sachant que nous ne nous intéressons pas aux échelles microscopiques par choix de ne pas considérer la mobilité quotidienne. On a ainsi schématiquement des échelles temporelles et spatiales mesoscopiques et des échelles macroscopiques.

Nous avons vu que la correspondance à des échelles temporelles et spatiales n'est pas systématique (voir la typologie provisoire à double entrée des processus). Par contre, celle à des domaines particuliers et à des acteurs l'est plus. Cette revue de littérature est donc orientée dans cette seconde direction.

### *Territoires*

Le courant principal s'intéressant à la modélisation de l'influence du réseau de transport sur les territoires se trouve dans le champ de la planification, à des échelles spatiales et temporelles moyennes (les échelles de l'accessibilité métropolitaine que nous avons développées ci-dessus). Des modèles en géographie à d'autres échelles, comme les modèles Simpop déjà évoqués [PUMAIN, 2012a], ne supposent pas une ontologie particulière pour le réseau de transport, et s'ils incluent des réseaux entre les villes comme porteur des échanges, ne permettent cependant pas d'étudier en particulier les relations entre réseaux et territoires. Nous reviendrons plus loin sur des extensions pertinentes pour notre question. Revoyons pour commencer un contexte de modèles plus proches des études de planification.

<sup>6</sup> Nous rappelons la signification de cette notation introduite au chapitre 1 : une flèche directe signifie des processus qu'on peut attribuer relativement de manière univoque à l'origine, tandis qu'une flèche réciproque suppose intrinsèquement l'existence d'interactions réciproques, généralement en coincidence avec l'émergence d'entités jouant un rôle dans celles-ci.

**MODÈLES LUTI** Ces approches sont désignées de manière générale comme *modèles d'interaction entre usage du sol et transport* (LUTI, pour *Land-Use Transport Interaction*). Il est entendu par usage du sol la répartition des activités territoriales, généralement réparties en typologies plus ou moins précises (par exemple logements, industrie, tertiaire, espace naturel). Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines scientifiques<sup>7</sup>. Leur principe général est de modéliser et simuler l'évolution de la distribution spatiale des activités, en prenant le réseau de transport comme contexte et déterminant significatif des localisations. Pour comprendre le cadre conceptuel sous-jacent à la majorité des travaux, l'Encadré 5 résume celui issu de [WEGENER et FÜRST, 2004]<sup>8</sup>.

Par exemple, du point de vue de l'économie urbaine, les propositions de tels modèles existent depuis un certain temps : [PUTMAN, 1975] rappelle le cadre d'économie urbaine où les principales composantes sont les emplois, la démographie et le transport, et passe en revue des modèles économiques de localisation qui s'apparentent au modèle de LOWRY déjà mentionné.

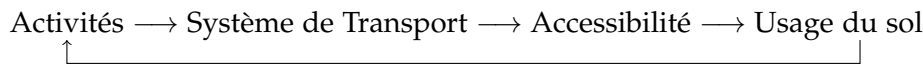
[WEGENER et FÜRST, 2004] donne plus récemment un état de l'art des études empiriques et de modélisation sur ce type d'approche des interactions entre usage du sol et transport. Le positionnement théorique est plutôt proche des disciplines de la socio-économie des transports et de la planification (voir les paysages disciplinaires dressés en 2.2). [WEGENER et FÜRST, 2004] compare et classe dix-sept modèles, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement courtes (de l'ordre de la décade) des simulations. On retrouve bien la correspondance avec les échelles typiquement mesoscopiques établies précédemment. Une revue complémentaire est faite par [CHANG, 2006], élargissant le contexte avec l'inclusion de classes plus générales de modèles, comme des modèles d'interactions spatiales (parmi lesquels l'attribution du trafic et les modèles à quatre temps), les modèles de planification basés sur la recherche opérationnelle (optimisation des localisations des différentes activités, généralement résidences et emplois), les modèles microscopiques d'utilité aléatoire, et les modèles de marché foncier.

Afin de donner une meilleure intuition de la logique sous-jacente à certains modèles Luti, nous détaillons en Encadré 6 et en Encadré 7 la

<sup>7</sup> Nous prenons le parti ici de rassembler de nombreuses approches ayant la caractéristique commune de modéliser principalement l'évolution de l'usage du sol, sur des échelles temporelles et spatiales moyennes. L'unité et le positionnement relatif de ces approches couvrant de l'économie à la planification, sont une question ouverte qui n'a à notre connaissance jamais été traitée de front. Le travail mené en 2.2 donne des pistes de réponse par une approche d'épistémologie quantitative.

<sup>8</sup> Un cadre plus général que nous avons déjà développé, qui permet de faire le pont avec notre cadre, est celui de [LE NÉCHET, 2010], qui remplace le triptyque Système de transport/Système de localisation/Système d'activités en relation avec les agents : agents demandeurs, agents aménageurs, facteurs externes.

[WEGENER et FÜRST, 2004] pose un cadre général théorique et empirique pour les modèles d'interaction entre transport et usage du sol. Les quatre concepts mobilisés sont l'usage du sol, la localisation des activités, le système de transport et la distribution de l'accessibilité. Un cycle d'effets circulaires sont résumés dans la boucle suivante :



Le système de transport est supposé à *infrastructure fixe*, c'est-à-dire que les effets de la distribution des activités sont ceux sur *l'utilisation* du système de transport (donc liés à la *mobilité* dans notre cadre plus général) : choix modal, fréquence des voyages, longueur des voyages.

Les effets théoriquement attendus sont classés selon les directions de la relation (*Usage du sol*→*Transport* ou *Transport*→*Usage du sol*, ainsi qu'une boucle *Transport*→*Transport*, qui fait partie des éléments ignorés dans notre cas), et par ailleurs par facteur agissant (densité résidentielle, d'emplois, localisation, accessibilité, coûts de transport) ainsi que par aspect affecté (longueur et fréquence des voyages, choix de mode, densités, localisations). On peut par exemple citer :

- *Usage du sol*→*Transport* : une densité résidentielle minimale est nécessaire pour l'efficienne du transport public, une concentration des emplois implique des voyages plus long, les villes plus grandes ont une part modale plus importante pour les transports en commun.
- *Transport*→*Usage du sol* : une forte accessibilité implique des prix plus élevés et un développement accru pour le résidentiel, les entreprises se localisent pour une meilleure accessibilité aux moyens de transport à grande échelle.
- *Transport*→*Transport* : les lieux avec une bonne accessibilité produiront plus et de plus longs voyages, le choix modal et le coût de transport sont fortement corrélés.

Ces effets théoriques sont par ailleurs comparés aux observations empiriques, qui pour la plupart donnent la manière d'implémentation du processus. Certains ne sont pas observés en pratique, tandis que la plupart sont en accord avec les attentes théoriques.

*Commentaire 1 : Un cadre uniscale ?* Ce cadre prend en compte schématiquement deux échelles principales, celle de la mobilité quotidienne et celle de la localisation des activités. Sachant qu'en pratique les comportement de mobilité sont généralement pris en compte sous forme de flux moyens, il se réduit souvent à une unique échelle mesoscopique. Dans tous les cas, il ne permet pas de tenir compte de dynamiques sur le temps plus long comprenant l'évolution de l'infrastructure du système de transport ou des dynamiques structurelles des systèmes de villes sur le temps long.

*Commentaire 2 : Une vision systématique des effets structurants ?* Par ailleurs, les pourfendeurs de la rhétorique des effets structurants trouveront en ce cadre leur bête noire, puisque les effets directs de l'accessibilité sur l'usage du sol puis sur la localisation des activités sont postulés ici. Ces critiques pourront être refoulées par l'observation qu'il s'agit des effets attendus théoriques, et que le cadre est mis en perspective des effets empiriques effectivement observés. Il sera à cependant à prendre avec précaution, en le situant toujours en terme de contexte et d'échelles.

**ENCADRÉ 5 : Cadre conceptuel des interactions entre transport et usage du sol selon [Wegener et Fürst, 2004].**



Le modèle Pirandello®<sup>a</sup> est présenté dans [DELONS, COULOMBEL et LEURENT, 2008] comme l'une des premières tentatives de développement de modèle Luti opérationnel en France. Le modèle se base sur quatre processus économiques fondamentaux : le marché du foncier et l'offre de logement, la mobilité résidentielle des ménages, l'attribution des destinations de déplacement, le choix modal. Le modèle est statique, c'est-à-dire calcule un équilibre pour les distributions spatiales des actifs et des emplois, ainsi que pour les flux de transport.

Les processus fondamentaux pris en compte et leur implémentation sont les suivants :

- Les choix résidentiels des ménages se basent sur une fonction d'utilité prenant en compte (i) un terme de confort en Cobb-Douglas de la surface et du revenu, corrigée par une préférence linéaire pour les logements individuels ; (ii) un terme d'accessibilité basé sur le coût généralisé (agrégation du coût de transport et du temps, avec un prix du temps) ; (iii) le prix du logement et de la taxe locale en fonction de la surface de logement ; (iv) un effet fixe par revenu et par zone ; et (v) un terme aléatoire supposé suivre une loi de Gumbel. Les probabilités de localisation pour une tranche de revenus sont alors données par un modèle de choix discret étant donné cette utilité.
- Les prix du logement sont formés selon une loi d'échelle de la population.
- Un système d'enchère local répond à la demande obtenue précédemment, en fonction d'une offre de logement exogène.
- Les entreprises se localisent en maximisant leur profit, fonction de la productivité (Cobb-Douglas du salaire et de l'accessibilité) et du prix du foncier, sous contrainte d'une distribution spatiale fixée du nombre d'emplois, de la surface de bureaux, et de la production totale de la région.
- Le transport est pris en compte par un modèle à quatre étapes, qui distribue les choix modaux et les choix de destination par un modèle de choix discrets, et les flux assignés selon un équilibre de Wardrop (voir ??), ce qui permet d'ajuster les valeurs de l'accessibilité étant donné une distribution spatiale des activités.

Le mécanisme de combinaison de ces différents processus pour obtenir un équilibre global est détaillé par [KRYVOKOV et al., 2013], et consiste à l'établissement de trois sous-équilibres à différentes échelles : flux de transport (donnant les coûts) sur le court terme, localisation et prix de l'immobilier sur le moyen terme, prix du foncier et terrain disponibles (fixés de manière exogène pour l'ensemble de la période modélisée).

*Commentaire : Équilibre, modèle opérationnel et calibration.* Un certain nombre de remarques peuvent être faites à ce modèle, les plus importantes pour notre approche étant : (i) l'hypothèse d'équilibre peut être un outil puissant pour comprendre la structure des attracteurs du système, mais n'a pas de fondement empirique, et encore moins pour le couplage d'équilibres à différentes échelles ; (ii) ainsi, la nature opérationnelle du modèle est discutable, puisque l'étude de l'impact de scénarios sur les déplacements des attracteurs permet difficilement d'inférer sur les dynamiques locales du système ; et (iii) les sous-modèles sont calibrés plus ou moins rigoureusement et relativement séparément, or les conditions d'un calibrage par décomposition sont une question ouverte encore peu explorée et liée à la nature du couplage de modèles. En notre sens, un tel modèle micro-fondé serait dans tous les cas en meilleure cohérence avec une philosophie de modélisation générative dynamique et de parcimonie (voir 3.1).

<sup>a</sup> L'origine du nom n'est pas donnée, mais suggère fortement l'influence de ses créateurs originaux V. PIRON et J. DELONS.



Le modèle Nedum2D, décrit en détail dans [VIGUIÉ, HALLEGATTE et ROZENBERG, 2014], se concentre sur la localisation des actifs et leur interaction avec la rente foncière et les promoteurs immobiliers : il s'agit d'un modèle inspiré du modèle de Fujita-Ogawa [FUJITA et OGAWA, 1982], héritant de la littérature en Economie Urbaine.

Les processus inclus dans le modèle sont, chacun ayant une échelle de temps particulière fixée par un paramètre :

- Les ménages font un compromis entre surface de logement et budget disponible hors coûts de transports et loyer, suivant une fonction de Cobb-Douglas pour l'utilité correspondante. Ce processus induit une dynamique pour la surface des logements en fonction de la distance au centre.
- Ils se relocalisent pour avoir une utilité moyenne plus grande que la moyenne.
- Les loyers évoluent pour maximiser l'occupation ou en réponse à une demande extérieure.
- De nouveaux bâtiments sont construits par des promoteurs cherchant à maximiser leur profits.

Ce modèle est dynamique et simule l'évolution de ces différentes variables dans l'espace (la formulation ci-dessous est monocentrique, une variante polycentrique et prenant en compte une distribution exogène d'emplois existent) et dans le temps. Son échelle spatiale est métropolitaine, et l'échelle de temps peut s'étendre d'une échelle moyenne (décade) à du temps plus long (siècle), sachant que cette dernière est peu crédible puisque qu'elle garde statique nombreuses autres composantes du système urbain.

*Commentaire : extension des ontologies.* Le couplage de Nedum avec un modèle d'attribution de transport, le modèle Modus<sup>a</sup>, vise à inclure la retroaction de la congestion dans le système de transport sur les coûts, et donc sur la localisation et la structure urbaine. Des questions fondamentales se dégagent des premières expériences de couplage :

- Le schéma directeur est-il vraiment utile, puisqu'il ne semble qu'accompagner des dynamiques déjà présentes. En d'autres termes, *le processus de gouvernance est-il endogène ?* Le Sdrif capture-t-il en fait une dynamique intrinsèque sur le temps plus long ?
- Le couplage des modèles pose en lui-même des difficultés techniques, de communication entre des modules déjà implémentés dans différents langages et de convergence du modèle couplé en un nombre raisonnable d'itérations.
- Il pose d'autre part des difficultés ontologiques : chaque modèle inclut des mécanismes opposés pour la même ontologie (effet d'agrégation contre congestion pour la distribution de la population). La question se pose s'il faut rajouter spécifiquement une ontologie de couplage (par exemple des equations spécifiques intégrant ces effets contradictoires), pour permettre d'une part une meilleure convergence, d'autre part une meilleur cohérence ontologique.

<sup>a</sup> Dans le cadre du projet en cours de réalisation ANR VITE! (voir <http://www.agence-nationale-recherche.fr/Projet-ANR-14-CE22-0013>).

structure, les ontologies et les hypothèses de deux modèles développés dans le cas spécifique de l'Ile-de-France (permettant d'une part la comparaison entre les deux et d'autre part donnant un écho au développement thématique de 1.2). Même pour des ontologies très proches (prix immobiliers, localisation des ménages), on voit la variété d'hypothèses possibles et de problématiques soulevée par les modèles.

**DES MODÈLES OPÉRATIONNELS TRÈS VARIÉS** La variété des modèles existants a conduit à des comparaisons opérationnelles : [PAULLEY et WEBSTER, 1991] rendent compte d'un projet comparant différents modèles appliqués à différentes villes. Leurs résultats permettent d'un part de classer des interventions en fonction de leur impact sur le niveau d'interaction entre transport et usage du sol, et d'autre part de montrer que l'effet des interventions dépend fortement de la taille de la ville et de ses caractéristiques socio-économiques.

Les ontologies des processus, et notamment sur la question de l'équilibre, sont aussi variées. Les avantages respectifs d'une approche statique (calcul d'un équilibre statique de la localisation des ménages pour une certaine spécification de leur fonctions d'utilité) et d'une approche dynamique (simulation hors équilibre des dynamiques résidentielles) a été étudié par [KRYVOBOKOV et al., 2013], dans un cadre métropolitain sur des échelles de temps de l'ordre de la décade. Les auteurs montrent que les résultats sont globalement comparables et que chaque modèle a son utilité selon la question posée.

Différents aspects du même système peuvent être traduits par divers modèles, comme le montre par exemple [WEGENER, MACKETT et SIMMONDS, 1991], et le trafic, les dynamiques résidentielles et d'emploi, l'évolution de l'usage du sol en découlant, influencée aussi par un réseau de transport statique, sont généralement pris en compte. [IACONO, LEVINSON et EL-GENEIDY, 2008] couvre un horizon similaire avec un développement supplémentaire sur les modèles à automates cellulaires d'évolution d'usage du sol et les modèles à base d'agents. La portée temporelle d'application de ces modèles, de l'ordre de la décade, et leur nature opérationnelle les rend utiles pour la planification, ce qui est assez loin de notre souci d'obtenir des modèles explicatifs de processus géographiques. En effet, il est souvent plus pertinent pour un modèle utilisé en planification d'être lisible comme outil d'anticipation, voire de communication, que d'être fidèle aux processus territoriaux au prix d'une abstraction.

**PERSPECTIVES POUR LES LUTI** [TIMMERMANS, 2003] émet des doutes quant à la possibilité de modèles d'interaction réellement intégrés, c'est-à-dire produisant des motifs de transports endogènes et se détachant d'artefacts comme l'accessibilité dont l'influence du caractère

artificiel reste à établir, notamment à cause du manque de données et une difficulté à modéliser les processus de gouvernance et de planification. Il est intéressant de noter que les priorités actuelles de développement des modèles LUTI semblent centrées sur une meilleure intégration des nouvelles technologies et une meilleure intégration avec la planification et les processus de prise de décision, par exemple via des interfaces de visualisation comme le propose [WEE, 2015]. Ils ne cherchent pas à s'étendre à des problématiques de dynamiques territoriales incluant le réseau sur de plus longues échelles par exemple, ce qui confirme la portée et la logique d'utilisation et de développement de ce type de modèles.

Une généralisation de ce type d'approche à une plus grande échelle, comme celle proposée par [RUSSO et MUSOLINO, 2012], consiste au couplage du LUTI à l'échelle mesoscopique à des modèles macroéconomiques à l'échelle macroscopique<sup>9</sup>. Ceux-ci ne considèrent pas l'évolution du réseau de transport de manière explicite mais s'intéressent seulement aux motifs abstraits d'offre et demande. L'économie urbaine a développé des approches spécifiques similaires dans leur démarche : [MASSON, 2000] décrit par exemple un modèle intégré couplant développement urbain, relocalisation et équilibre des flux de transports.

Ainsi, nous pouvons synthétiser ce type d'approche, qu'on pourra désigner par abus de langage *approche LUTI*, par les caractéristiques fondamentales suivantes : (i) modèles visant à comprendre une évolution du territoire, dans le contexte d'un réseau de transport donné ; (ii) modèles dans une logique de planification et d'applicabilité, étant souvent impliqués eux-même dans les prises de décision ; et (iii) modèles à des échelles moyennes, dans l'espace (métropole) et dans le temps (décade).

#### *Croissance du Réseau*

Passons à présent au paradigme "opposé", centré sur l'évolution du réseau. Il peut sembler incongru de considérer un réseau variable en négligeant les variations du territoire, au regard de l'aperçu de certains des mécanismes potentiels d'évolution revus précédemment (rupture de potentiel, auto-renforcements, planification du réseau) qui se produisent à des échelles de temps majoritairement plus longues que les évolutions territoriales. On verra ici qu'il n'y a pas de paradoxe, vu que (i) soit la modélisation s'intéresse à l'évolution des *propriétés du réseau*, à une courte échelle (micro) pour des processus de

9 [RUSSO et MUSOLINO, 2012] généralise en fait le cadre des LUTI pour proposer un cadre d'interaction entre Economie Spatiale et Transports (*Spatial Economics and Transport Interactions*). Celui-ci inclut les LUTI à l'échelle urbaine, et au niveau national les modèles macroéconomiques simulant production et consommation, compétition des activités, production du stock d'offre de transport. Les modèles de transport supposent toujours réseau fixe et établissent des équilibres au sein de celui-ci, ce qui implique une petite échelle spatiale et une courte échelle temporelle.

congestion, de capacité, de tarification, principalement d'un point de vue économique ; (ii) soit les composantes territoriales jouant en effet sur le réseau sont stables aux échelles longues considérées.

La croissance de réseaux est l'objet de démarches de modélisation qui cherchent à expliquer la croissance des réseaux de transport. Ils prennent généralement un point de vue *bottom-up* et endogène, c'est-à-dire cherchant à mettre en évidence des règles locales qui permettraient de reproduire la croissance du réseau sur de longues échelles de temps (souvent le réseau routier). Comme nous allons le voir, il peut s'agir de la croissance topologique (création de nouveaux liens) ou la croissance des capacités des liens en relation avec leur utilisation, selon les échelles et les ontologies considérées. Nous distinguons pour simplifier des grands courants disciplinaires s'étant intéressés à la modélisation de la croissance des réseaux de transport : ceux-ci sont liés respectivement à l'économie des transports, la physique, la géographie des transports et la biologie.

On rejoint ainsi partiellement la classification de [XIE et LEVINSON, 2009c], qui propose une revue étendue de la modélisation de croissance des réseaux, dans une perspective d'économie des transports mais en élargissant à d'autres champs. [XIE et LEVINSON, 2009c] distingue des grands courants disciplinaires ayant étudiés la croissance des réseaux de transport : la géographie des transports a développé très tôt des modèles basés sur des faits empiriques mais qui ont visé à reproduire la topologie plutôt que sur les mécanismes<sup>10</sup> ; les modèles statistiques sur des cas d'étude fournissent des conclusions très mitigées sur les relations causales entre croissance du réseau et demande (la croissance étant dans ce cas conditionnée aux données de demande) ; les économistes ont étudié la production d'infrastructure à la fois d'un point de vue microscopique et macroscopique, généralement non spatialisés ; la science des réseaux a produit des modèles stylisés de croissance de réseau qui se basent sur des règles topologiques et structurelles plutôt que des règles se reposant sur des processus correspondant à des réalités empiriques.

**ECONOMIE** Les économistes ont proposé des modèles de ce type : [ZHANG et LEVINSON, 2007] passe en revue la littérature en économie des transports sur la croissance des réseaux, rappelant les trois aspects principalement traités par les économistes sur le sujet, qui sont la tarification routière, l'investissement en infrastructures et le régime de propriété, et propose finalement un modèle analytique combinant les trois. Ces trois classes de processus relèvent d'une interaction entre les agents économiques microscopiques (utilisateurs du réseau) et les agents de gouvernance. Les modèles peuvent inclure

<sup>10</sup> Selon [XIE et LEVINSON, 2009c], la contribution de la géographie consiste en des efforts limités à l'époque de [HAGGETT et CHORLEY, 1970], nous nous baserons donc sur cette revue et ne donnerons pas de développement approfondi.

une description détaillée des processus de planification, comme [LEVINSON, XIE et OCA, 2012] qui combine des enquêtes qualitatives et des statistiques pour paramétrer un modèle de croissance de réseau. [XIE et LEVINSON, 2009b] compare l'influence relative des processus de croissance centralisés (planification par une structure de gouvernance) et décentralisés (croissance locale ne rentrant pas dans le cadre d'une planification globale). [LEVINSON et KARAMALAPUTI, 2003] procède à une étude empirique des déterminants de la croissance du réseau routier pour les *Twin Cities* aux Etats-Unis (Minneapolis-Saint-Paul), établissant que les variables basiques (longueur, changement dans l'accessibilité) ont le comportement attendu, et qu'il existe une différence entre les niveaux d'investissement, impliquant que la croissance locale n'est pas affectée par les coûts, ce qui peut correspondre à une équité des territoires en termes d'accessibilité. Ces données sont utilisées par [ZHANG et LEVINSON, 2017] pour calibrer un modèle de croissance de réseau qui superpose les décisions d'investissement aux motifs d'utilisation du réseau. [YERRA et LEVINSON, 2005] montre avec un modèle économique basé sur des processus d'auto-renforcement (c'est-à-dire incluant une rétroaction positive des flux sur la capacité) et incluant une règle d'investissement basée sur l'attribution du trafic, que des règles locales sont suffisantes pour faire émerger une hiérarchie du réseau routier à usage du sol fixé. Une synthèse de ces travaux gravitant autour de LEVINSON est faite dans [XIE et LEVINSON, 2011a].

**PHYSIQUE** La physique a introduit récemment des modèles de croissance des réseaux d'infrastructure, en s'inspirant largement de cette littérature économique : un modèle très similaire au dernier cité est donné par [LOUF, JENSEN et BARTHELEMY, 2013] avec des fonctions coûts-bénéfices plus simples mais obtenant une conclusion similaire. Étant donné une distribution de noeuds (villes)<sup>11</sup> dont la population suit une loi puissance, deux villes seront connectées par un lien routier si une fonction d'utilité coût-bénéfice, combinant linéairement flux gravitaire potentiel et coût de construction<sup>12</sup>, a une valeur positive. Ces hypothèses locales simples suffisent à faire émerger un réseau complexe et des transitions de phase en fonction du paramètre de poids relatif dans le coût, conduisant à l'apparition de la hiérarchie. [ZHAO et al., 2016] applique ce modèle de manière itérative pour connecter des zones intra-urbaines, et montre que la prise en compte

<sup>11</sup> Nous nous trouvons ici dans un cas où l'hypothèse de non-évolution des population des villes tandis que le réseau s'établit itérativement trouve peu de support empirique ou théorique, puisqu'on a montré que réseau et villes avaient des échelles de temps d'évolution comparables. Ce modèle produit donc plus à proprement parler un *réseau potentiel* étant donné une distribution de villes, et il est à interpréter avec précaution.

<sup>12</sup> Ce qui donne une fonction de coût de la forme  $C = \beta/d_{ij}^\alpha - d_{ij}$ , où  $\alpha$  et  $\beta$  sont des paramètres

des populations dans la fonction de coût change significativement les topologies obtenues.

Une autre classe de modèles, proche dans leur idée des modèles procéduraux, se basent sur des processus d'optimisation géométrique locale, et visent à ressembler à des réseaux réels dans leur topologie. [BOTTINELLI, LOUF et GHERARDI, 2017] étudie ainsi un modèle de croissance d'arbre appliqué aux pistes de fourmis, dans lequel coût de maintenance et coût de construction influencent tous les deux les choix de nouveau lien. Le modèle de morphogenèse de [COURTAT, GLOAGUEN et DOUADY, 2011] qui utilise un compromis entre réalisation des potentiels d'interaction et coût de construction, ainsi que des règles de connectivité, reproduit de manière stylisée des motifs réels des réseaux de rues. Un modèle très proche est décrit dans [RUI et al., 2013], tout en incluant des règles supplémentaires pour l'optimisation locale (prise en compte du degré pour la connection de nouveaux liens). La conception optimale de réseau, plutôt pratiquée par l'ingénierie, utilise des paradigmes similaires : [VITINS et AXHAUSEN, 2010] explore l'influence de différentes règles d'une grammaire de formes (notamment les motifs de connection entre les liens de différents niveaux hiérarchiques) sur les performances de réseaux générés par algorithme génétique.

Détaillons les mécanisme de l'un de ces modèles de croissance géométrique. [BARTHELEMY et FLAMMINI, 2008] décrit un modèle basé sur une optimisation locale de l'énergie qui génère des réseaux routiers à l'aspect globalement crédible. Le modèle suppose des "centres", qui correspondent à des noeuds d'un réseau routier, et des segments de route dans l'espace reliant ces centres. Le modèle part de centre initiaux connectés, et procède par itérations pour simuler la croissance du réseau de la façon suivante :

1. De nouveaux centres sont ajoutés aléatoirement suivant une distribution de probabilité exogène, aux pas de temps multiple d'une durée fixée.
2. Le réseau croît suivant une règle de minimisation de coût : les centres sont groupés par projection sur le réseau ; chaque groupe fait croître un segment de longueur fixée dans la direction moyenne vers l'ensemble du groupe à partir de la projection (sauf si celle-ci est nulle, un segment croît alors en direction de chaque point).

Ce modèle est ajusté pour que les aires des parcelles délimitées par le réseau suivent une loi d'échelle avec un exposant similaire à celui observé pour la ville de Dresde. Il a l'avantage d'être simple, d'avoir peu de paramètres (distribution de probabilité pour les centres, taille des tronçons construits), de reposer sur des règles locales crédibles. Cette dernière propriété est à double tranchant, puisqu'on peut alors s'attendre à ce que le modèle ne puisse capturer que peu de complexité,



en négligeant de nombreux processus mis en valeur au chapitre 1 comme la gouvernance.

**RÉSEAUX BIOLOGIQUES** Enfin, une approche originale et intéressante pour la croissance des réseaux est le réseau biologique. Cette approche appartient au champ de l'ingénierie morphogénétique, qui vise à concevoir des systèmes complexes artificiels inspirés de systèmes complexes naturels et sur lesquels un contrôle des propriétés émergentes est possible [DOURSAT, SAYAMA et MICHEL, 2012]. Les *machines Physarum*, qui sont des modèles d'une moisissure auto-organisée (*slime mould*) ont été prouvés comme résolvant de manière efficace des problèmes difficiles (au sens de leur complexité computationnelle, voir 3.3) comme des problèmes de routage [TERO, KOBAYASHI et NAKAGAKI, 2006] ou des problèmes de navigation NP-complets comme le Problème du Voyageur de Commerce [ZHU et al., 2013a]. Ces propriétés permettent à ces systèmes de produire des réseaux ayant des propriétés de coût-robustesse Pareto-efficaces [TERO et al., 2010] qui sont typiques des propriétés empiriques des réseaux réels, et de plus relativement proches en forme de ceux-ci (sous certaines conditions, voir [ADAMATZKY et JONES, 2010]).

Ce type de modèles peut être d'intérêt dans notre cas puisque les processus d'auto-renforcement basés sur les flots sont analogues aux mécanismes de renforcement de lien en économie des transports. Ce type d'heuristique a été testé pour générer le réseau ferré Français par [MIMEUR, 2016], faisant un pont intéressant avec les modèles d'investissement de LEVINSON décrits précédemment<sup>13</sup>.

**MODÉLISATION PROCÉDURALE** Finalement, nous pouvons mentionner d'autres tentatives comme [DE LEON, FELSEN et WILENSKY, 2007; YAMINS, RASMUSSEN et FOGEL, 2003], qui sont plus proches de la modélisation procédurale [LECHNER et al., 2004; WATSON et al., 2008] et pour cette raison n'ont que peu d'intérêt pour notre cas puisqu'ils peuvent difficilement être utilisés comme modèles explicatifs<sup>14</sup>. La modélisation procédurale consiste à générer des structures à la ma-

<sup>13</sup> Sachant que pour cette étude, les critères de validation appliqués restent toutefois limités, soit à un niveau inadapté aux faits stylisés étudiés (nombre d'intersection ou de branches) soit trop généraux et pouvant être produit par n'importe quel modèle (longueur totale et pourcentage de population desservie), et relèvent de critères de forme typique de la modélisation procédurale qui ne peuvent que difficilement rendre compte des dynamiques internes d'un système comme développé précédemment. De plus, prendre pour validation externe la production d'un réseau hiérarchique découle d'une exploration incomplète de la structure et du comportement du modèle, puisque celui-ci par ses mécanismes d'attachement préférentiel doit mécaniquement produire une hiérarchie. Ainsi, une attention particulière devra être donnée au choix des critères de validation.

<sup>14</sup> Suivant [VARENNE, 2017], un modèle explicatif permet de produire une explication à des régularités ou des lois observées, par exemple en suggérant des processus pouvant en être à l'origine. Si les processus du modèle sont explicitement dissociés d'une ontologie raisonnable, ceux-ci ne peuvent être explications potentielles. Nous

nière des grammaires de forme<sup>15</sup>, mais celle-ci se concentre généralement sur la reproduction fidèle de forme locale, sans tenir compte des propriétés macroscopiques émergentes. Les classer comme modèles de morphogenèse n'est pas correct et correspond à une incompréhension des mécanismes du *Pattern Oriented Modeling* [GRIMM et al., 2005]<sup>16</sup> d'une part et de l'épistémologie de la morphogenèse d'autre part (voir 5.1). Nous utiliserons ce type de modèle (mélange d'exponentielles pour produire une densité de population par exemple) pour générer des données synthétiques initiales uniquement pour paramétrer d'autres modèles complexes (voir 3.1 et 5.3).

### 2.1.3 Modéliser la co-évolution

Nous pouvons à présent nous intéresser aux modèles intégrant dynamiquement le paradigme Territoire ↔ Réseau, qui on le rappelle suppose qu'un conditionnement de l'un par l'autre n'est pas identifiable. Les ontologies utilisées, comme nous le verrons, couplent<sup>17</sup> souvent des éléments de réseau avec des composantes territoriales, mais cette position n'est pas une nécessité et certains éléments peuvent être hybrides (par exemple une structure de gouvernance du système de transport peut relever simultanément des deux aspects). Dans notre lecture des modèles, ces différentes spécifications se dégageront naturellement.

Nous désignerons largement par modèle de co-évolution les modèles de simulation qui incluent un couplage des dynamiques de la croissance urbaine et du réseau de transport. Ceux-ci sont relativement rares, et pour la plupart au stade de modèles stylisés. Les efforts étant assez disparates et dans des domaines très variés, il y a peu d'unité dans ces approches, si ce n'est l'abstraction de l'hypothèse d'interdépendance entre réseaux et caractéristiques du territoire dans le temps. Nous proposons de les passer en revue toujours avec la grille de lecture des échelles.

---

donnerons en 3.1 un développement de cette notion dans le cadre d'une réflexion plus générale sur l'épistémologie de la modélisation.

- 15 Une grammaire de forme est un système formel (c'est-à-dire un ensemble de symboles initiaux, les axiomes, et un ensemble de règles de transformation) qui agit sur des objets géométriques. Partant de motifs initiaux, elles permettent de générer des classes d'objets
- 16 Le *Pattern Oriented Modeling* consiste à chercher à expliquer des motifs observés, généralement à plusieurs échelles, dans une démarche *bottom-up*. La modélisation procédurale n'en relève pas, puisqu'elle vise à reproduire et non à expliquer.
- 17 Nous rappelons la définition du couplage de modèle, qui correspond à celle de couplage de système ou de processus donnée en introduction : il s'agit de la constitution d'un modèle qui est simultanément une extension de chacun des modèles initiaux.

*Echelle microscopique et mesoscopique*

**MODÈLES GÉOMÉTRIQUES** [ACHIBET et al., 2014] décrit un modèle de co-évolution à une très grande échelle (échelle du bâtiment), dans lequel l'évolution du réseau et des bâtiments sont tous les deux régis par un agent commun, influencé différemment par la topologie du réseau et la densité de population, qui peut être compris comme un agent développeur. Le modèle permet de simuler une extension urbaine auto-organisée et de produire des configurations de quartier. Bien qu'il couple fortement composantes territoriales (bâtiments) et réseau routier, les résultats présentés ne permettent pas de tirer de conclusion sur les processus de co-évolution en eux-mêmes.

Une généralisation du modèle d'optimisation locale géométrique décrit précédemment est développée dans [BARTHELEMY et FLAMMINI, 2009], et cherche à capturer la co-évolution entre topologie du réseau et densité de ses noeuds. La localisation de nouveaux noeuds est influencée à la fois par la densité et la centralité, permettant de boucler le couplage fort. Plus précisément, le fonctionnement global du modèle est le même, ainsi que la règle de croissance du réseau. Les centres se localisent quant à eux selon une fonction d'utilité qui est une combinaison linéaire de la centralité de chemin moyenne dans un voisinage et de l'opposée de la densité (dispersion due aux prix plus élevés en fonction de la densité). Cette utilité permet de définir la probabilité de localisation des nouveaux centres suivant un modèle de choix discrets. Le modèle permet de montrer que l'influence de la centralité accentue les phénomènes d'agrégation (notamment par une résolution analytique sur une version en une dimension du modèle), et reproduit par ailleurs des profils exponentiels décroissants pour la densité (loi de Clarke), observés empiriquement.

[DING et al., 2017] introduisent un modèle de co-évolution entre différentes couches du réseau de transport, et montrent l'existence d'un paramètre de couplage optimal en terme d'inégalités de centralité pour la conception d'un réseau : si on assimile le réseau routier à granularité très fine à une distribution de population, ce modèle se rapproche du précédent modèle de co-évolution entre réseau de transport et territoire.

**MODÈLES ÉCONOMIQUES** [LEVINSON, XIE et ZHU, 2007] prennent une approche économique plus riche du point de vue des processus de développement de réseau impliqués, similaire à un modèle à quatre étapes (c'est-à-dire incluant une génération de flux origine-destination et une attribution du trafic dans le réseau) qui inclut coût de transport et congestion, couplé avec un module d'investissement routier qui simule les revenus des péages pour les agents qui construisent, et un module d'évolution d'usage du sol qui simule les relocalisations des actifs et des emplois. Les expériences d'exploration de ce modèle montrent que l'usage du sol et le réseau en co-

évolution conduisent à des retroactions positives renforçant les hiérarchies. Elles sont cependant loin d'être satisfaisantes puisque la topologie du réseau n'évolue pas à proprement parler puisque seules les capacités et les flux changent dans le réseau, ce qui signifie que des mécanismes plus complexes (comme la planification de nouvelles infrastructures) sur de plus longues échelles de temps ne sont pas pris en compte. [LI et al., 2016] ont récemment étendu ce modèle par l'ajout de prix immobiliers endogènes et d'une heuristique d'optimisation par algorithme génétique pour les agents décideurs.

D'un autre point de vue, [LEVINSON et CHEN, 2005] est aussi présenté comme un modèle de co-évolution mais repose sur un modèle prédictif à chaîne de Markov, et donc plus proche d'une analyse statistique que d'un modèle de simulation basé sur des processus. [RUI et BAN, 2011] décrivent un modèle dans lequel le couplage entre usage du sol et la topologie du réseau est fait par un paradigme faible, l'usage du sol et l'accessibilité n'ayant pas de retroaction sur la topologie du réseau, le modèle d'usage du sol étant conditionné à la croissance du réseau autonome.

**AUTOMATES CELLULAIRES** Un modèle hybride simple exploré et appliqué à un exemple stylisé de planification de la répartition fonctionnelle d'un nouveau quartier dans [RAIMBAULT, BANOS et DOURSAT, 2014], repose sur les mécanismes d'accès aux activités urbaines pour la croissance des établissements avec un réseau s'adaptant à la forme urbaine. Les règles pour la croissance du réseau sont trop simples pour capturer des processus plus élaborés qu'une simple connection systématique (comme une rupture de potentiel par exemple), mais le modèle produit à une grande échelle une large gamme de formes urbaines qui reproduisent les motifs typiques des établissements humains. Ce modèle s'inspire de [MORENO, BADARIOTTI et BANOS, 2012] pour ses mécanismes de base mais permet une génération de formes bien plus larges par la prise en compte des fonctions urbaines.

A ces échelles relativement grandes, s'étendant de l'échelle urbaine à celle métropolitaine, les mécanismes de localisation de population influencée par l'accessibilité couplés à des mécanismes de croissance de réseau optimisant certaines fonctions semblent être la règle pour ces modèles : de la même façon, [WU et al., 2017] couplent un automate cellulaire de diffusion de population à un réseau optimisant un coût local dépendant de la géométrie et de la distribution de population.

Des modèles répondant à des problématiques assez lointaines peuvent par ailleurs être reliés à notre question : par exemple, de manière conceptuelle, une certaine forme de couplage fort est également utilisé dans [BIGOTTE et al., 2010] qui par une approche de recherche opérationnelle proposent un algorithme de conception de réseau pour

optimiser l'accessibilité aux services, prenant en compte à la fois la hiérarchie du réseau et celle des centres connectés.

Ainsi, les modèles de co-évolution aux échelles microscopique et mesoscopiques suivent globalement la structure suivante : (i) processus de localisation ou relocalisation des activités (actifs, bâtiments) influencés par leur propre distribution et les caractéristiques du réseau ; (ii) évolution du réseau, topologique ou non, répondant à des règles très diverses : optimisation locale, règles fixes, planification par des agents décideurs. Cette diversité suggère la nécessité de prendre en compte la superposition de multiples processus régissant l'évolution du réseau.

### *Modélisation de Systèmes Urbains*

A une échelle macroscopique, la co-évolution est parfois prise en compte dans des modèles de systèmes urbains. [BAPTISTE, 1999] propose de coupler un modèle de croissance urbaine basé sur les migrations (introduit par l'application de la synergetique aux systèmes de villes par [SANDERS, 1992]) avec un mécanisme d'auto-renforcement des capacités pour le réseau routier sans modification topologique. Plus précisément, les principes généraux du modèle sont les suivants.

- Des indicateurs d'attractivité et de répulsion permettent pour chaque ville de déterminer des taux d'émigration et d'immigration et de faire évoluer les populations.
- La topologie du réseau est fixée dans le temps, mais les capacités des liens évoluent. La règle est une augmentation de la capacité lorsque le flux dépasse celle-ci par un seuil donné comme paramètre pendant un certain nombre d'itérations. Les flux sont affectés par un modèle gravitaire d'interaction entre les villes.

Sa dernière version est présentée par [BAPTISTE, 2010]. Les conclusions générales qui peuvent être tirées de ce travail sont que ce couplage permet de faire émerger une configuration hiérarchique<sup>18</sup> et que l'ajout du réseau produit un espace moins hiérarchique, permettant à des villes moyennes de bénéficier de la rétroaction du réseau de transport.

Le modèle proposé par [BLUMENFELD-LIEBERTHAL et PORTUGALI, 2010] peut être vu comme un pont entre l'échelle mesoscopique et les approches de type système urbain, puisqu'il simule les migrations

<sup>18</sup> Mais on sait par ailleurs que des modèles plus simples, un attachement préférentiel uniquement par exemple, permettent de reproduire ce fait stylisé. Le modèle doit avoir pour objectif de répondre à des problématiques plus larges, comme la compréhension fine des processus de co-évolution, ce qui n'est pas fait ici. Cependant, l'un de ses objectifs opérationnels est par ailleurs rempli, par l'application à la France et l'étude de l'impact d'un projet de Ligne à Grande Vitesse, rappelant les multiples fonctions possibles d'un modèle (voir 3.1).

entre villes et la croissance du réseau induite par une rupture de potentiel lorsque les détours sont trop grands. Dans la continuité des modèles Simpop pour modéliser les systèmes de villes, [SCHMITT, 2014] décrit le modèle SimpopNet qui vise à précisément intégrer les processus de co-évolution dans les systèmes de villes à longue échelle temporelle, typiquement par des règles pour un développement hiérarchique du réseau comme fonction des dynamiques des villes, couplées à celles-ci qui dépendent de la topologie du réseau. Malheureusement le modèle n'a pas été exploré ni étudié de manière plus approfondie, et de plus est resté au niveau de modèle jouet. [COTTINEAU, 2014] propose une croissance endogène des réseaux de transport comme la dernière brique de construction du cadre de modélisation Marius, mais cela reste à un niveau conceptuel puisque cette brique n'a pas encore été spécifiée ni implémentée. Il n'existe à notre connaissance pas de modèle empirique ou appliqué à un cas concret se basant sur une approche de la co-évolution par les systèmes urbains vus par la théorie évolutive des villes.

On voit bien l'opposition aux principes épistémologiques de l'économie géographique : [FUJITA, KRUGMAN et MORI, 1999] introduisent par exemple un modèle évolutionnaire capable de reproduire une hiérarchie urbaine et une organisation typique de la théorie des places centrales [BANOS et al., 2011], mais qui repose toujours sur la notion d'équilibres successifs, et surtout considère un modèle "à-la-Krugman" c'est-à-dire un espace à une dimension, isotrope, et dans lequel les agents sont répartis de manière homogène<sup>19</sup>. Cette approche peut être instructive sur les processus économiques en eux-mêmes mais plus difficilement sur les processus géographiques, puisque ceux-ci impliquent un déroulement des processus économiques dans l'espace géographique dont les particularités spatiales qui ne sont pas prises en compte dans cette approche sont essentielles. Notre travail s'attachera à montrer dans quelle mesure cette structure de l'espace peut être importante et également explicative, puisque les réseaux, et encore plus les réseaux physiques induisent des processus spatio-temporels dépendants au chemin et donc sensibles aux singularités locales et propices aux bifurcations induites par la combinaison de celles-ci et de processus à d'autres échelles (par exemple la centralité induisant un flux).

A l'échelle macroscopique, les modèles existants se basent sur des évolutions des agents (souvent les villes) en conséquence de leurs interactions, portées par le réseau, tandis que l'évolution du réseau peut répondre à différentes règles : auto-renforcement, rupture de potentiel. La structure générale est globalement la même qu'à des

<sup>19</sup> L'absence d'espace réel n'est pas un problème dans cette approche économique qui vise à comprendre des processus hors-sol. Dans notre cas, la structure de l'espace géographique n'est pas séparable, et même au coeur des problématiques qui nous intéressent.



TABLE 4 : **Synthèse des approches de modélisation.** Le type donne le sens de la relation ; la classe est le champ scientifique dans lequel le modèle se place ; les échelles correspondent à nos échelles simplifiées ; les fonctions sont données au sens de 3.1 ; nous donnons enfin le type de résultat qu'ils fournissent et les paradigmes utilisés.

Type	Classe	Echelle Tem- porelle	Echelle Spa- tiale	Fonction	Résultats	Paradigmes
Réseaux → Territoires	LUTI	Moyenne	Mesoscopique	Planification, Prédiction	Simulation de l'usage du sol	Économie urbaine
Territoires →	Économie des Ré- seaux	Moyenne	Mesoscopique	Explication	Rôle de pro- cessus éco- nomiques	Économie, Gouver- nance
Réseaux	Croissance géomé- trique	Longue	Meso ou Ma- cro	Explication	Reproduction de formes stylisées	Modèles de Simulation, Optimisa- tion locale
	Réseaux biologiques	Longue	Mesoscopique	Optimisation	Production de réseaux optimaux	Réseau auto- organisé
Territoires ↔	Économie des Ré- seaux	Moyenne	Mesoscopique	Explication	Effets de ren- forcement	Économie
Réseaux	Croissance géomé- trique	Longue ou NA	Micro, Meso ou Macro	Explication	Reproduction de formes stylisées	Modèles de Simulation, Optimisa- tion locale
	Systèmes Urbains	Moyenne, Longue	Macroscopique	Explication, prospection	Faits stylisés	Géographie complexe

échelles plus grandes, mais les ontologies restent fondamentalement différentes.

### Synthèse

Il est essentiel à ce stade de s'oser à une synthèse et une mise en perspective de l'ensemble des modèles que nous avons passé en revue, puisque même si celle-ci sera nécessairement réductrice et simplificatrice, elle donne les fondations pour les analyses qui suivront.

Nous synthétisons les grands types de modèles que nous avons passé en revue dans le tableau suivant, en les classant par type (relation entre réseaux et territoires), par classe (grandes classes correspondant à la stratification de la revue), et en précisant les échelles temporelle et spatiales concernées, les fonctions, le type de résultats obtenus, les paradigmes utilisés. Celle-ci est donnée en Table 4.

*Une co-évolution négligée ?*

Le déséquilibre entre la dernière section rendant compte des modèles intégrant effectivement une dynamique fortement couplée (et possiblement une co-évolution) et les précédentes interroge : les modèles intégrant la co-évolution sont-ils si marginaux ? Est-il alors possible d'expliquer cette marginalité ?

L'objet des deux sections qui suivent sera de proposer des éléments de réponse à ces questions par des analyses épistémologiques en accroissant la connaissance des champs concernés et des modèles correspondants.



Nous avons ainsi donné dans cette section un aperçu large des modèles s'intéressant aux interactions entre réseaux de transport et territoires, incluant les modèles de co-évolution. Nous commençons donc à entrevoir une précision de la définition du concept de co-évolution dans ce cadre.

Nous proposons dans la section suivante de dresser une cartographie plus systématique de ce paysage scientifique, afin de renforcer le point de vue épistémologique et mieux situer la position que nous prendrons et les modèles que nous introduirons par la suite.



## 2.2 UNE APPROCHE ÉPISTÉMOLOGIQUE

Nous avons eu un aperçu large de différents types de modèles prenant en compte les interactions entre réseaux et territoires, ainsi que les disciplines et problématiques associées. Ces aspects très différents suggèrent un cloisonnement fort des disciplines. Il reste de plus difficile de situer les modèles potentiels de co-évolution dans cette nébuleuse. Il est légitime de se demander quelles sont les relations existantes et potentielles entre les différentes approches ? Quelles domaines peuvent être passés inaperçus bien que complémentaires ?

Diverses hypothèses peuvent être avancées pour tenter d'expliquer l'absence d'investigation des modèles de co-évolution :

- Suivant [COMMENGES, 2013], les acteurs scientifiques et opérationnels qui seraient concernés par l'application pratique de tels modèles se verraient remplacés par ces mêmes modèles et donc n'ont aucune incitation à les développer (explication sociologique).
- Les différentes disciplines qui détiennent les diverses composantes nécessaires à de tels modèles sont cloisonnées et ont des motivations divergentes (explication épistémologique).
- La construction de tels modèles comporte des difficultés intrinsèques rendant leur développement décourageant et pas parfaitement maîtrisé actuellement.

Nous n'aurons pas les moyens d'explorer la première hypothèse (ou plutôt elle demanderait un sujet à part entière, impliquant entre autres entretiens sociologiques). La troisième est soit une tautologie soit indémontrable, à-la-Church dirait-on, et l'ensemble de notre travail permettra d'y apporter des pistes de réponse. La deuxième par contre est comme nous allons le voir plus à notre portée.

Une manière d'explorer cette hypothèse et de répondre aux questions précédentes consiste en une étude épistémologique que nous proposons de mener de manière quantitative et systématique. Cette approche est complémentaire de l'analyse de littérature précédente, et permet à la fois de la contextualiser et de la systématiser. Il faut par ailleurs garder en tête l'idée que l'étude des raisons de la rareté des modèles nous informera nécessairement sur les modèles eux-mêmes et les questions reliées à leur construction : la *connaissance de la connaissance* [MORIN, 1986] accroît la connaissance.

Une étude préliminaire a pour but de confirmer la pertinence d'une approche d'épistémologie quantitative, en suggérant une forte isolation des disciplines. Celle-ci est menée par un algorithme de revue systématique algorithmique, qui reconstruit des corpus de références par exploration de voisinage sémantiques, c'est-à-dire la récupération itérative de références voisines dans leur contenu sémantique principal. Nous procédons ensuite à une analyse de réseaux, couplant

réseau de citation et réseau sémantique, pour préciser les contours des disciplines impliquées. Nous suggérons finalement des possibles extensions vers de l'apprentissage non-supervisé et la fouille de texte complets pour une extraction automatique de la structure de modèles par exemple.

Commençons par situer le contexte des analyses en *épistémologie quantitative*<sup>20</sup> que nous proposons de mener.

### 2.2.1 *Epistémologie quantitative*

Les méthodes possibles pour des entrées quantitatives en épistémologie sont nombreuses. Une bonne illustration de la variété des approches est donnée par l'analyse de réseau. En utilisant des caractéristiques topologiques du réseau de citations, un bon pouvoir prédictif pour les motifs de citation est par exemple obtenu par [NEWMAN, 2014]. Les réseaux de co-auteurs peuvent également être utilisés pour des modèles prédictifs [SARIGÖL et al., 2014]. Une approche par réseau multi-couches est proposée dans [OMODEI, DE DOMENICO et ARENAS, 2017], qui utilise des réseaux bipartites d'articles et de chercheurs, afin de produire des mesures d'interdisciplinarité en utilisant des mesures de centralité généralisées. Les disciplines peuvent être stratifiées en couches pour révéler des communautés entre celles-ci et ainsi des motifs de collaboration [BATTISTON et al., 2016]. Les réseaux de mots-clés sont utilisés dans d'autres champs comme en économie de l'innovation : par exemple, [CHOI et HWANG, 2014] propose une méthode pour identifier les opportunités technologiques par la détection de mots-clés importants du point de vue des mesures topologiques. De façon similaire, [SHIBATA et al., 2008] utilise une analyse topologique du réseau de citations pour détecter des fronts de recherche émergents.

### *Revue systématique*

Avec l'avènement des nouveaux moyens techniques et des nouvelles sources de données, la revue de littérature classique tend à se coupler à des revues automatiques. Des techniques de revue systématique ont été développées, des revues qualitatives aux méta-analyses quantitatives qui permettent de produire des nouveaux résultats par combinaison d'études existantes [RUCKER, 2012]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'information

<sup>20</sup> Nous proposons d'utiliser ce terme pour des travaux à la croisée de la bibliométrie et de la scientométrie, des sciences cognitives, de l'épistémologie et des systèmes complexes, à l'image de l'*Epistémologie Appliquée* développée jusqu'en 2011 par le laboratoire CREA.

qui par l'accès plus aisé à l'information rend difficilement justifiable l'omission de références clés [LISSACK, 2013]<sup>21</sup>.

### *Interdisciplinarité*

Le développement d'approches interdisciplinaires est de plus en plus nécessaire pour la plupart des disciplines, à la fois pour la découverte de nouvelles connaissances mais aussi pour l'impact sociétal des découvertes, comme le rappelle récemment le volume spécial de la revue *Nature* (NATURE, 2015). [BANOS, 2013] suggère que leur développement doit s'insérer dans une spirale subtile entre et au sein des disciplines. Une autre façon de voir ce phénomène est de le comprendre comme l'émergence de champs verticalement intégrés<sup>22</sup> de manière conjointe aux questions horizontales comme détaillé dans la feuille de route des systèmes complexes [BOURGINE, CHAVALARIAS et AL., 2009].

Il existe naturellement de multiples points de vue sur ce qu'est exactement l'interdisciplinarité (de nombreux d'autres termes comme la trans-disciplinarité ou la cross-disciplinarité existent aussi) et cela dépend en fait des domaines impliqués : des disciplines hybrides apparues récemment (voir par exemples celles soulignées par [BAIS, 2010] comme l'astro-biologie, ou d'autres plus proche de notre champ comme la géomatique) sont une bonne illustration du cas où les intrications sont très fortes, tandis que des champs comme "l'urbanisme" dont les définitions sont multiples montrent dans quelle mesure l'intégration horizontale est nécessaire et comment de la connaissance transversale peut être produite. Les interactions entre les disciplines ne sont pas toujours faciles, comme le montrent les malentendus lorsque les sujets sur la ville ont été récemment introduits aux physiciens comme [DUPUY et BENGUIGUI, 2015] le rappelle, malentendus dont les effets peuvent être négatifs s'ils conduisent à des conflits ou à une ignorance de connaissances déjà établies par un autre domaine.

Ces questions font partie de la compréhension des processus de production de connaissance, i.e. la *connaissance de la connaissance* comme [MORIN, 1986] la présente, dans laquelle les perspectives *evidence-based*, qui impliquent des approches quantitatives, jouent un rôle important. Ces paradigmes peuvent être compris comme une *épistémologie quantitative*. Des mesures quantitatives de l'interdisciplinarité feraient pour cette raison partie d'une approche multidimensionnelle de l'étude de la science, qui va en quelque sorte "au delà de la biblio-

21 Tout en restant conscient que même avec une méthode systématique, il est impossible d'être absolument exhaustif. L'objectif est d'augmenter autant que possible la couverture, dans l'idée d'une approche inclusive de multiples points de vue, comme le propose notre positionnement épistémologique de perspectivisme donné en 3.3.

22 C'est-à-dire intégrant, généralement entre les échelles, différentes branches d'un champ : par exemple la biologie intégrative [LIU, 2005] vise à des ponts entre approches génomiques, approches physiologiques, approches écologiques, en tirant parti de l'intégration des méthodes : expérimentation, modélisation, simulation.

métrie” [CRONIN et SUGIMOTO, 2014]. La préoccupation de cette section se positionne dans ce champ de recherche. Nous passons d’abord en revue les approches existantes à la mesure de l’interdisciplinarité.

Les définitions de l’interdisciplinarité elle-même et les indicateurs pour la mesurer ont déjà été traités par un vaste corpus de littérature. [HUUTONIEMI et al., 2010] rappelle la différence entre les approches *multi-disciplinaires* (une agrégation de travaux de différentes disciplines) et *interdisciplinaires* (qui implique un certain niveau d’intégration). Ils construisent un cadre qualitatif pour classifier différents types d’interdisciplinarité, et distinguent par exemple les interdisciplinarités empiriques, théoriques et méthodologique. L’aspect multi-dimensionnel de l’interdisciplinarité est confirmé même au sein d’un champ spécifique comme la littérature [AUSTIN et al., 1996]. Une première façon de quantifier l’interdisciplinarité d’un ensemble de publications est de regarder la proportion de disciplines hors d’une discipline principale dans lesquelles elles sont publiées, comme [RINIA, LEEUWEN et RAAN, 2002] fait pour l’évaluation de projets en physique, de manière complémentaire au jugement d’experts. [PORTER et al., 2007] désigne cette mesure comme *spécialisation*, et la compare avec une mesure d’intégration donnée par l’étendue des citations faites par un article au sein des différentes *Subject Categories* (classification du *Web of Knowledge*), qui est également appelé indice de *Rao-Stirling*. [LARIVIÈRE et GINGRAS, 2010] l’utilise sur un corpus du *Web of Science* pour montrer l’existence d’un niveau d’interdisciplinarité intermédiaire optimal pour l’impact en termes de citations sur une fenêtre de 5 ans post-publication. Un travail équivalent est fait dans [LARIVIÈRE et GINGRAS, 2014], qui se concentre sur l’évolution des mesures sur une longue portée temporelle. L’influence des données manquantes sur cet index est étudié par [MORENO, AUZINGER et WERTHNER, 2016], qui fournit un cadre étendu qui prend en compte l’incertitude. L’utilisation de réseaux a également été proposée : [PORTER et RAFOLS, 2009] combine l’indice d’intégration avec une technique de cartographie qui consiste en la visualisation de réseaux synthétiques construits par les co-citations entre disciplines. [LEYDESDORFF, 2007] montre que la centralité de chemin est un indicateur pertinent d’interdisciplinarité, lorsqu’un environnement de citation pertinent est considéré.

### 2.2.2 Revue Systématique Algorithmique

Nous proposons de procéder de manière préliminaire à une revue de la littérature systématique et algorithmique. Un algorithme itératif formel pour construire des corpus de références à partir de mots-clés initiaux, basé sur l’analyse textuelle, est développé et mis en oeuvre. Nous étudions ses propriétés de convergence et procédons à une analyse de sensibilité. Nous l’appliquons ensuite à des requêtes repré-



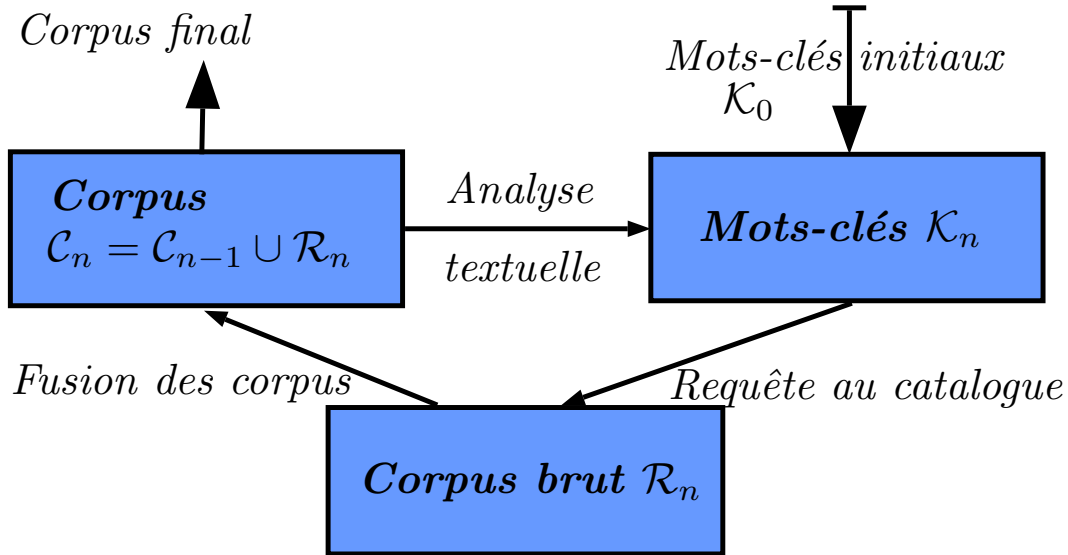


FIGURE 7 : **Architecture globale de l'algorithme.** A partir d'un ensemble de mots-clés initiaux, on construit un corpus par requête au catalogue, duquel on extrait des nouveaux mots-clés par analyse textuelle. On itère alors en boucle jusqu'à obtenir un corpus fixe ou dépasser un nombre maximal fixé d'itérations.

sentatives de notre question spécifique, pour lesquelles les résultats tendent à confirmer l'hypothèse d'isolation relative des disciplines.

Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citations [NEWMAN, 2014] ou les réseaux de co-auteurs [SARIGÖL et al., 2014], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [CHAVALARIAS et COINTET, 2013], qui produit une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. Nous prenons le parti d'une appréhension de la diversité des domaines, introduite en 2.1, par cette information supplémentaire du paysage scientifique. Les méthodes que nous introduisons sont particulièrement adaptées pour notre étude puisque nous voulons comprendre la structure du contenu des recherches sur le sujet.

L'algorithme procède par itérations pour obtenir un corpus stabilisé à partir de mots-clés initiaux, reconstruisant l'horizon sémantique scientifique autour d'un sujet donné. La description formelle de l'algorithme est détaillée en Annexe A.2, avec les détails de son implémentation et des analyses de sensibilité. Sa logique est donnée par le schéma en Fig. 7 : étant donné un ensemble de mots-clés de départ que l'on rassemble en une unique requête, on récolte des travaux qui en traitent, dont on extrait de nouveaux mots-clés pour itérer en boucle jusqu'à convergence éventuelle.

Nous partons de cinq différentes requêtes initiales qui ont été manuellement extraites des divers domaines identifiés dans la biblio-

graphie<sup>23</sup>, afin de comparer les corpus obtenus pour chaque requête. Après avoir construit les corpus, nous étudions leur cohérence lexicale comme un indicateur de réponse à notre question initiale. De grande distances devraient confirmer l’hypothèse formulée ci-dessus, i.e. que des disciplines auto-centrées pourraient être à l’origine d’un manque d’intérêt pour des modèles co-évolutifs. La table 5 montre les valeurs de la proximité lexicale relative, que nous définissons par un indice de similarité d’ensemble pondéré donné par

$$d(I, J) = \frac{\sum_{k_i \in I, k_j \in J} \mathbb{1}_{k_i=k_j} \cdot (s(k_i) + s(k_j))}{\sum_{k_i \in I} s(k_i) + \sum_{k_j \in J} s(k_j)}$$

pour les corpus  $I, J$ , et avec  $s$  fonction strictement positive donnant une mesure d’importance des mots au sein des corpus fournie par la méthode d’extraction des mots-clés (voir A.2). Ses valeurs sont significativement faibles en comparaison à la valeur de référence 1 pour des corpus égaux (la mesure s’interprète comme une proportion de mots en co-occurrence), ce qui tend à confirmer notre hypothèse<sup>24</sup>.

La constatation d’un faible nombre de modèles qui simulent la co-évolution des réseaux de transport et de l’usage du sol urbain pourrait être due à l’absence de communication entre les disciplines scientifiques étudiant différents aspects du problème. D’autres explications possibles qui en sont proches peuvent par exemple être le manque de cas d’application concrets de tels modèles vu les échelles temporelles mises en jeu et donc l’absence de financement propre - ce qui n’est pas si loin de l’absence d’une discipline y consacrant certains de ses objets. Cette question des portées et des échelles des modèles fera l’objet de la meta-analyse à la section suivante 2.3. Ainsi, nous avons proposé ici une méthode algorithmique pour donner des éléments de réponse par l’extraction de corpus basée sur l’analyse textuelle, dont les résultats numériques semblent aller dans le sens d’une compartimentation des disciplines (au sens particulier utilisé ici d’une distance sémantique entre corpus niches). Cette analyse était relativement limitée dans la portée de ses résultats, notamment par le

<sup>23</sup> Qui sont “cityANDsystemANDnetwork”, “land-useANDtransportANDinteraction”, “networkANDurbanANDmodeling”, “populationANDdensityANDtransport”, “transportationANDnetworkANDurbanANDgrowth”. Ce choix inclut les approches par systèmes de villes, les approches LUTI, les approches de croissance de réseau. Il ne peut bien sûr être exhaustif. Cette étude étant préliminaire on admet de travailler potentiellement sur des échantillons. Par exemple, l’utilisation de “co-evolution” n’est pas concluante car trop peu d’articles utilisent cette formulation. De même, la question de la langue conditionne les résultats : une requête en Français conduit à des niches linguistiques finalement assez pauvres en diversité, et nous faisons ainsi uniquement des requêtes en anglais. L’approche par hyper-réseau développée plus loin sera elle multilingue.

<sup>24</sup> Pour situer ces résultats de manière relative, il faudrait un modèle nul (c’est-à-dire générateur de corpus avec distributions sémantiques similaires mais sans structure de corrélation entre mots) avec des corpus aléatoires par exemple, ce qui pourrait faire l’objet de développements futurs.

TABLE 5 : Matrice symétrique des proximités lexicales entre les corpus finaux, définies comme la somme des co-occurrences pondérées de mots-clés finaux entre corpus, normalisé par le poids total des mots-clés finaux. La taille des corpus finaux est donnée par  $W$ . Les valeurs obtenues pour les proximités sont considérablement faibles par rapport à la valeur maximale 1, ce qui confirme que les corpus sont éloignés de manière significative.

Corpus	1	2	3	4	5
1 ( $W=3789$ )	1	0	0.0719	0.0078	0.0724
2 ( $W=5180$ )	0	1	0.0338	0	0.0125
3 ( $W=3757$ )	0.0719	0.0338	1	0.0100	0.1729
4 ( $W=3551$ )	0.0078	0	0.0100	1	0.0333
5 ( $W=8338$ )	0.0724	0.0125	0.1729	0.0333	1

faible nombre de requêtes et un certain nombre d'incertitudes intrinsèques, mais est *suffisante* pour produire un diagnostique, à savoir (i) une structure disciplinaire fortement marquée peut être extraite de l'analyse de corpus, et (ii) l'utilisation d'outils sémantique permet une extraction d'information endogène. Fort de ce diagnostique préliminaire, nous proposons d'approfondir l'analyse par une variation et extension de la méthode employée.

### 2.2.3 Bibliométrie indirecte

Comme décrit précédemment, l'analyse sémantique des corpus finaux ne contient pas la totalité de l'information sur les liens entre disciplines ni sur les motifs de propagation de la connaissance scientifique comme ceux contenus dans les réseaux de citations par exemple. De plus, la collection des données dans l'algorithme précédent est sujette à convergence vers des thèmes relativement auto-cohérents de par la structure propre de la méthode. Il serait possible d'obtenir plus d'information sur les motifs sociaux de choix ontologiques pour la modélisation en étudiant les communautés dans des réseaux plus larges, ce qui correspondrait plus à des disciplines (ou des sous-disciplines selon le niveau de granularité). Nous proposons de reconstruire les disciplines autour de notre thématique, pour obtenir une vue plus précise du paysage scientifique sur notre sujet et des liens entre disciplines. Une contribution fondamentale de cette section consiste en la construction de jeux de données hybrides à partir de sources hétérogènes, et les développement des outils associés qui peuvent être réutilisés et améliorés pour des applications similaires.

Cette démarche peut être vue comme une bibliométrie indirecte<sup>25</sup>, puisqu'on cherche à reconstruire une information endogène et à extraire des relations entre différentes dimensions.

### *Contexte*

L'approche développée ici couple exploration et analyse de réseau de citations avec analyse textuelle, dans le but de cartographier le paysage scientifique dans le voisinage d'un corpus donné. Le contexte est particulièrement intéressant pour la méthodologie développée. Premièrement, le sujet étudié est très large et par essence interdisciplinaire. Deuxièmement, les données bibliographiques sont difficiles à obtenir, soulevant la question de comment la perception d'un horizon scientifique peut être déterminée par les acteurs de la dissémination et donc loin d'être objective, rendant les solutions techniques comme celle développée ici en conséquence des outils cruciaux pour une science ouverte et neutre.

Notre approche combine une analyse des communautés sémantiques (comme fait dans [PALCHYKOV et al., 2016] pour les articles en physique mais sans extraction des mots-clés, ou par [GURCIULLO et al., 2015] pour un analyse des réseaux sémantiques de débats politiques) avec celle du réseau de citations pour extraire par exemple des mesures d'interdisciplinarité. Cette contribution se démarque des travaux précédents quantifiant l'interdisciplinarité puisqu'elle ne suppose pas de domaines a priori ou une classification des références considérées, mais reconstruit par le bas les champs via l'information sémantique endogène. [NICHOLS, 2014] introduit une approche similaire, utilisant le modèle d'extraction de thématiques *Latent Dirichlet Allocation*<sup>26</sup> pour caractériser l'interdisciplinarité de récompenses dans des sciences précises.

### *Données*

Notre approche implique des spécifications pour le jeu de données utilisé, à savoir : (i) couvrir un voisinage conséquent du corpus étudié dans le réseau de citation afin d'avoir une vue la moins biaisée possible du paysage scientifique ; (ii) avoir au moins une description textuelle pour chaque noeud. Pour cela, nous rassemblons et compilons les données de sources hétérogènes en utilisant une architecture et

<sup>25</sup> La bibliométrie, ou scientométrie lorsqu'elle est appliquée en particulier à la science comme dans notre cas, consiste en la mesure et la qualification des motifs de production de connaissance par l'intermédiaire de leur supports directement observables (productions scientifiques, fonctionnement des institutions, relations sociales entre chercheurs, etc.) [CRONIN et SUGIMOTO, 2014]. Cet ouvrage rappelle que ce domaine est en pleine mutation et dresse une carte des nouvelles approches.

<sup>26</sup> Le modèle LDA, introduit par [BLEI, NG et JORDAN, 2003], suppose les documents comme produits par des thèmes sous-jacents, avec une distribution de Dirichlet pour leur composition ainsi que pour la distribution des mots par thèmes. Son estimation donne la composition des thèmes en terme de mots-clés.

implémentation spécifiques, décrites en Annexe B.6. Pour simplifier, nous dénommons *référence* toute production scientifique standard<sup>27</sup> qui peut être citée par une autre (articles de journaux, livre, chapitre de livre, article d’actes, communication, etc.) et contient des informations de base (titre, résumé, auteurs, année de publication). Nous travaillons par la suite sur le réseau des références.

**CORPUS INITIAL** Notre corpus initial est construit à partir de l’état de l’art établi en 2.1. Sa composition complète est donnée en Annexe A.2. Il s’agit de 7 références “phares” identifiées pour chacune des disciplines abordées précédemment. Le but ici n’est pas d’être exhaustif (cela le sera en 2.3), mais de construire une description du voisinage des domaines qui nous concernent. Celui-ci est pris de taille raisonnable (conduisant à un réseau final traitable sans méthode spécifique concernant la taille des données), mais les méthodes utilisées ici ont été développées sur des données massives, pour les brevets par exemple [BERGEAUD, POTIRON et RAIMBAULT, 2017a], et comme il le sera en Annexe F à l’ensemble de notre bibliographie.

**DONNÉES DE CITATION** Le réseau de citations est reconstruit à partir de Google Scholar qui est souvent l’unique source des citations entrantes [NORUZI, 2005] puisqu’en science humaines les ouvrages ne sont pas systématiquement référencés par les bases fournissant des services (payants) comme le réseau de citation.<sup>28</sup> Nous sommes conscient des biais possibles de l’utilisation de cette source unique (voir par exemple [BOHANNON, 2014])<sup>29</sup>, mais ces critiques sont dirigées vers les résultats de recherche plutôt que les comptes de citations. Nous récoltons ainsi les références *citantes* à profondeur deux, c’est-à-dire les références citant le corpus initial et celles citant celles-ci. Le réseau obtenu contient  $V = 9462$  références correspondant à  $E = 12004$  liens de citation. Concernant les langues, l’anglais représente 87% du corpus, le français 6%, l’espagnol 3%, l’allemand 1%, complété par des langues comme le mandarin pouvant être indéfinies (la détection de celui-ci étant peu fiable).

**DONNÉES TEXTUELLES** Pour mener l’analyse sémantique, une description suffisamment conséquente est nécessaire. Nous collectons pour cela les résumés pour le réseau précédent. Ceux-ci sont disponibles pour environ un tiers des références, donnant  $V = 3510$  noeuds avec description textuelle.

27 Ce qui est bien sûr sujet à débat, voir nos discussions en ouverture sur l’évolution des modes de communication scientifique.

28 Par exemple, le journal Cybergeos n’est indexé dans le *Web of Science* que depuis mai 2016, suite à des négociations ardues et non sans contrepartie.

29 Ou <http://iscpif.fr/blog/2016/02/the-strange-arithmetic-of-google-scholars>.

TABLE 6 : Description et taille des communautés de citation.

Domaine	Taille (% de noeuds)
LUTI	18%
Géographie Urbaine et des Transports	16%
Planification des infrastructures	12%
Planification intégrée - TOD	6%
Réseaux Spatiaux	17%
Etudes d'accessibilité	18%

### Résultats

**RÉSEAU DE CITATIONS** Des statistiques basiques pour le réseau de citation donnent déjà des informations intéressantes. Le réseau a un degré moyen de  $\bar{d} = 2.53$  et une densité de  $\gamma = 0.0013^{30}$ . Le degré entrant moyen (qui peut être interprété comme un facteur d'impact stationnaire) est de 1.26, ce qui est relativement élevé pour des sciences humaines. Il est important de noter sa connexité faible, ce qui signifie que les domaines initiaux ne sont pas en isolation totale : les références initiales sont partagées à un degré minimal par les différents domaines. Nous travaillons sur la suite sur le sous-réseau des noeuds comprenant au moins deux liens, pour extraire le coeur de la structure du réseau et se débarrasser de l'effet "grappe". De plus, le réseau est nécessairement complet entre ces noeuds puisqu'on est remonté au deuxième niveau.

Nous procédons pour le réseau de citation à une détection de communautés par l'algorithme de Louvain, sur le réseau non-dirigé correspondant. L'algorithme fournit 13 communautés, de modularité dirigée 0.66<sup>31</sup>, extrêmement significative en comparaison à une estimation par bootstrap de la même mesure sur le graphe aléatoirement rebranché qui donne une modularité de  $0.0005 \pm 0.0051$  sur  $N = 100$  répétitions. Les communautés font sens de manière thématique, puisqu'on retrouve pour les plus grosses les domaines présentés dans la Table 6.

Les appellations sont à regard d'expert a posteriori, selon les grands domaines dégagés dans la revue de littérature en 2.1<sup>32</sup>.

<sup>30</sup> Pour référence, [BATAGELJ, 2003] présente les caractéristiques de 11 réseaux scientifiques de domaines divers et de taille allant de 40 à 8851 noeuds, et reporte des densité variant de  $3.3 \cdot 10^{-4}$  à 0.038, avec une médiane à 0.003, proche de celle de notre réseau.

<sup>31</sup> La modularité est une mesure du "niveau de clustering" d'une partition d'un réseau en classes. L'algorithme de Louvain construit les communautés par optimisation gourmande de la modularité.

<sup>32</sup> On note que cette dénomination est bien exogène et nécessairement subjective. Comme développé plus loin pour le réseau sémantique, il n'existe pas de technique



La Fig. 8 montre le réseau de citation et permet de visualiser les relations entre ces domaines. Il est intéressant d’observer que les travaux des économistes et des physiciens dans le domaine tombent dans la même catégorie d’étude des *Spatial Networks*. En effet, la littérature citée par les physiciens comporte souvent plus d’ouvrage en économie qu’en géographie, tandis que les économistes utilisent des techniques d’analyse de réseau. Ensuite, le planning, l’accessibilité, les LUTI et le TOD sont très proches mais se distinguent dans leur spécificités : le fait qu’ils apparaissent dans des communautés séparées témoigne d’un certain niveau de cloisonnement. Ceux-ci font le pont entre les approches réseaux spatiaux et les approches géographiques, qui comportent une partie importante de sciences politiques par exemple. Les liens entre physique et géographie restent très faibles. Ce panorama dépend bien sûr du corpus initial, mais nous permet de mieux comprendre le contexte de celui-ci dans son environnement disciplinaire.

**COMMUNAUTÉS SÉMANTIQUES** L’extraction des mots-clés est faite suivant une heuristique inspirée de [CHAVALARIAS et COINTET, 2013]. La description complète de la méthode et de son implémentation est donnée en Annexe B.6. Elle se base sur les relations au second ordre entre les entités sémantiques, qui sont des *n-grams*, c’est-à-dire des mots-clés multiples pouvant avoir une longueur jusqu’à 3. Celles-ci sont estimées via la matrice de co-occurrence, dont les propriétés statistiques fournissent une mesure de déviation à des co-occurrences uniformes, qui est utilisée pour juger la pertinence des mots-clés. Sélectionnant un nombre fixe de mots-clés pertinents  $K_W = 10000$ , nous pouvons ensuite construire un réseau pondéré par les co-occurrences.

La topologie du réseau brut ne permet pas l’extraction claire de communautés, en particulier à cause de hubs qui correspondent à des termes fréquents commun à de nombreux champs (e.g. `model`, `space`). Ces mots sont utilisés de manière comparable dans l’ensemble des champs étudiés, et ne portent pas d’information pour les séparer<sup>33</sup>. Nous faisons l’hypothèse que ces termes à fort degré ne portent pas d’information particulière sur des classes données et peuvent ainsi être filtrés étant donné un seuil de degré maximal  $k_{\max}$  (on s’intéresse alors à ce qui fait la spécificité de chaque domaine). De la même manière, les liens avec un poids faibles sont considérés comme du bruit et filtrés selon un seuil de poids minimal  $\theta_w$ . La méthode générique permet de plus une filtration préliminaire des mot-clés, complémentaire à la filtration topologique, par fréquence d’apparition dans les documents  $[f_{\min}, f_{\max}]$ , à laquelle les résultats ne sont pas sensibles dans notre cas. L’analyse de sensibilité des caractéristiques du

simple pour une désignation endogène. Il faut garder cet aspect en tête pour la mise en perspective des interprétations et conclusions.

<sup>33</sup> Mais en porteraient si l’on comparait un corpus de géographie quantitative et un corpus de musicologie par exemple.

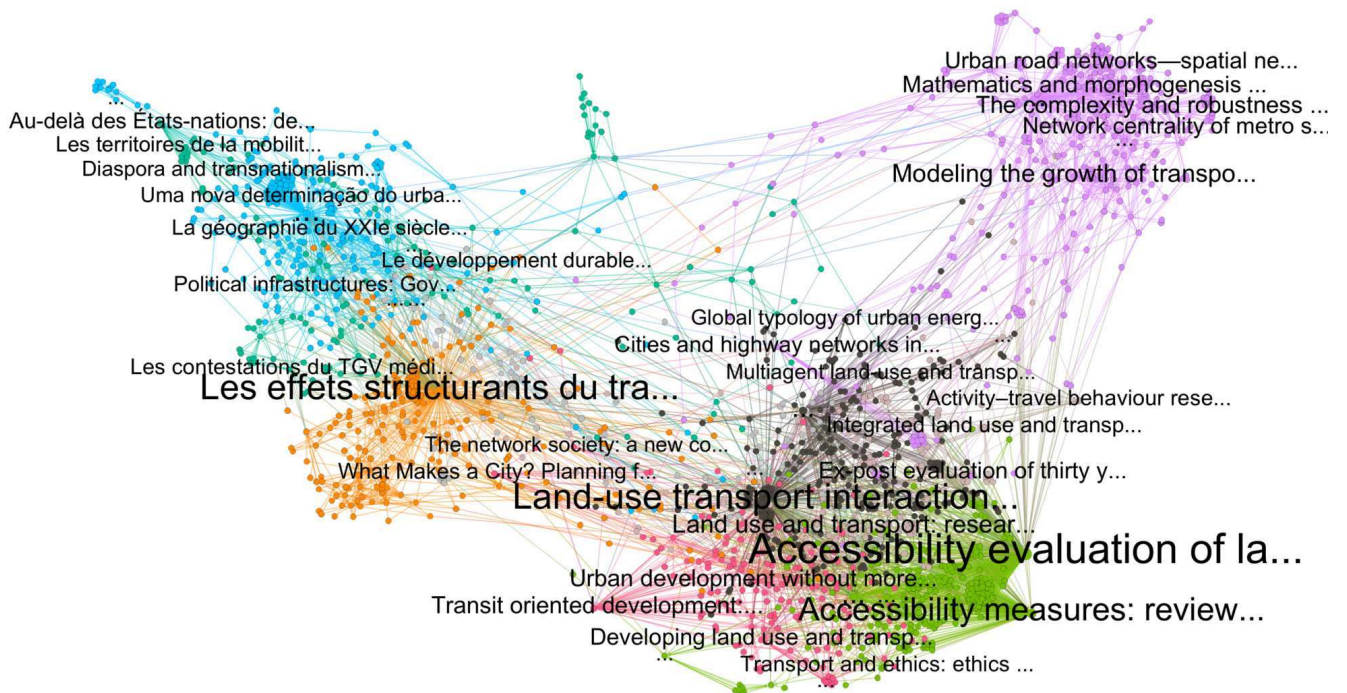


FIGURE 8 : **Réseau de citations.** Nous visualisons les références ayant au moins deux liens, par un algorithme de force-atlas. Les couleurs donnent les communautés décrites dans le texte. En orange, bleu, turquoise : géographie urbaine, géographie des transports, sciences politiques ; en rose, noir, vert : planning, accessibilité, LUTI ; en violet : réseaux spatiaux (physique et économie).

réseau filtré, notamment de sa taille, modularité et structure des communautés, est donnée en A.2. Nous choisissons des valeurs de paramètres permettant une optimisation multi-objectifs entre modularité et taille du réseau,  $\theta_w = 10$ ,  $k_{max} = 500$ , par le choix d'un point compromis sur un front de Pareto, qui donne un réseau sémantique de taille ( $V = 7063$ ,  $E = 48952$ ). Celui-ci est visualisé en Annexe A.2.

Nous récupérons ensuite les communautés dans le réseau par un clustering de Louvain standard sur le réseau filtré optimal. On obtient 20 communautés pour une modularité de 0.58. Celles-ci sont examinées à la main pour être nommées, les techniques de nomination automatique [YANG et al., 2000] ne sont pas assez élaborées pour faire la distinction implicite entre champs thématiques et méthodologiques par exemple (en fait entre les domaines de connaissance, voir 8.3) qui est une dimension supplémentaire que nous ne traitons pas ici, mais nécessaire pour avoir des descriptions parlantes. Les communautés

sont décrites en Table 7. On voit tout de suite la complémentarité avec l'approche par citations, puisque se dégagent ici à la fois des sujet d'étude (High Speed Rail, Maritime Networks), des domaines et méthodes (Networks, Remote Sensing, Mobility Data Mining), des domaines thématiques (Policy), des méthodes pures (Agent-based Modeling, Measuring). Ainsi, une référence peut mobiliser plusieurs de ces communautés. On a de plus une granularité plus fine de l'information. L'effet du langage est puissant puisque la géographie française se distingue en une catégorie séparée (des analyses poussées pourraient être envisagées pour mieux comprendre le phénomène et en tirer parti : sous-communautés, reconstruction d'un réseau spécifique, études par traduction ; mais celles-ci sont hors de propos dans cette étude exploratoire). On constate l'importance des réseaux, des problématiques de sciences politiques et socio-économiques. Nous mobiliserons la première catégorie dans la plupart des modèles développés, mais en gardant en tête l'importance des problématiques liées à la gouvernance, nous réaliserons un travail spécifique en 7.3.

**MESURES D'INTERDISCIPLINARITÉ** La distribution des mots clés dans les communautés permettent de définir une mesure d'interdisciplinarité au niveau de l'article. La combinaison des couches de citation et sémantique dans l'hyperréseau fournit des mesures d'interdisciplinarité au second ordre (motifs sémantiques des cités ou des citants), que nous n'utiliserons pas ici à cause de la taille modeste du réseau de citation (voir B.6 et C.5). Plus précisément, une référence  $i$  peut être vue comme un vecteur de probabilités sur les classes sémantiques  $j$ , qu'on notera sous forme matricielle  $\mathbf{P} = (p_{ij})$ . Celles-ci sont estimées simplement par les proportions de mots-clés classifiés dans chaque classe pour la référence. Une mesure classique d'interdisciplinarité [BERGEAUD, POTIRON et RAIMBAULT, 2017a] est alors  $I_i = 1 - \sum_j p_{ij}^2$ . Soit  $\mathbf{A}$  la matrice d'adjacence du réseau de citation, et soit  $\mathbf{I}_k$  les matrices de selection des lignes correspondants à la classe  $k$  de la classification de citation :  $\text{Id} \cdot \mathbb{1}_{c(i)=k}$ , telle que  $\mathbf{I}_k \cdot \mathbf{A} \cdot \mathbf{I}_{k'}$  donne exactement les citations de  $k$  vers  $k'$ . La proximité de citation entre les communautés de citation est alors définie par  $c_{kk'} = \sum \mathbf{I}_k \cdot \mathbf{A} \cdot \mathbf{I}_{k'} / \sum \mathbf{I}_k \cdot \mathbf{A}$ . On définit la proximité sémantique en définissant une matrice de distance entre références par  $\mathbf{D} = d_{ii'} = \sqrt{\frac{1}{2} \sum (p_{ij} - p_{i'j})^2}$  puis la proximité sémantique par  $s_{kk'} = \mathbf{I}_k \cdot \mathbf{D} \cdot \mathbf{I}_{k'} / \sum \mathbf{I}_k \sum \mathbf{I}_{k'}$ .

Nous montrons en Fig. 9 les valeurs de ces différentes mesures, ainsi que la composition sémantique des communautés de citation, pour les classes sémantiques majoritaires. La distribution de  $I_i$  montre que les articles gravitant dans le domaine du LUTI sont les plus interdisciplinaires dans les termes utilisés, ce qui pourrait être lié à leur caractère appliqué. Les autres disciplines sont dans des motifs similaires, à part la géographie et la planification des infrastructures qui

TABLE 7 : **Description des communautés sémantiques.** On donne leur taille, leur proportion en quantité de mots-clés (sous la forme de *multi-stems*) cumulés sur l'ensemble du corpus, et des mots-clés représentatifs sélectionnés par degré maximal.

Name	Size	Weight	Keywords
Networks	820	13.57%	social network, spatial network, resili
Policy	700	11.8%	actor, decision-mak, societi
Socio-economic	793	11.6%	neighborhood, incom, live
High Speed Rail	476	7.14%	high-spe, corridor, hsr
French Geography	210	6.08%	système, développement, territoire
Education	374	5.43%	school, student, collabor
Climate Change	411	5.42%	mitig, carbon, consumpt
Remote Sensing	405	4.65%	classif, detect, cover
Sustainable Transport	370	4.38%	sustain urban, travel demand, activity-bas
Traffic	368	4.23%	traffic congest, cbd, capit
Maritime Networks	402	4.2%	govern model, seaport, port author
Environment	289	3.79%	ecosystem servic, regul, settlement
Accessibility	260	3.23%	access measur, transport access, urban growth
Agent-based Modeling	192	3.18%	agent-bas, spread, heterogen
Transportation planning	192	3.18%	transport project, option, cba
Mobility Data Mining	168	2.49%	human mobil, movement, mobil phone
Health Geography	196	2.49%	healthcar, inequ, exclus
Freight and Logistics	239	2.06%	freight transport, citi logist, modal
Spanish Geography	106	1.26%	movilidad urbana, criteria, para
Measuring	166	1.0%	score, sampl, metric

présentent des distributions quasi-uniformes, témoignant de l'existence de références très spécialisées dans ces classes. Ce n'est pas nécessairement étonnant vu les sous-champs pointus exhibés (sciences politiques par exemples, et de même les études prospectives type coût-bénéfices sont très étriquées). Ce premier croisement des couches nous confirme les spécificités de chaque champ. Concernant les compositions sémantiques, la plupart agissent comme validation externe vu les classes majoritaires. Le champ le moins concerné par les problèmes socio-économiques est la planification des infrastructures, ce qui donnera du grain à moudre aux détracteurs de la technocratie. Les questions de changement climatique et durabilité sont relativement bien réparties. Enfin, les ouvrages géographiques concernent en majorité des problèmes de gouvernance.

Les matrices de proximité confirment la conclusion obtenue précédemment en termes de citation, les partages étant très faibles, les plus hautes valeurs étant jusqu'à un quart de la planification vers la géographie et des LUTI vers le TOD (mais pas l'inverse, les relations peuvent être à sens unique). Or les proximités sémantiques montrent par exemple que LUTI, TOD, Accessibility et Networks sont proches dans leur termes, ce qui est logique pour les trois premiers, et confirme pour le dernier que les physiciens se basent majoritairement sur les méthodes de ces champs liés au planing pour légitimer leur travaux. La géographie est totalement isolée, sa plus proche voisine étant la planification des infrastructures. Cette étude est très utile pour notre propos, puisqu'elle montre des domaines cloisonnés partageant des termes et donc a priori des problématiques et sujets communs. Les domaines ne se parlent pas tout en parlant des langages pas si lointains, d'où la pertinence accrue de vouloir accorder leur partitions dans nos travaux : nos modèles devront mobiliser des éléments, ontologies et échelles de ces différents champs.

Nous concluons cette analyse par une approche plus robuste pour quantifier les proximités entre couches de l'hyperréseau. Il est aisé de construire une matrice de corrélation entre deux classifications, par les corrélations de leur colonnes. Nous définissons les probabilités  $P_C$  toutes égales à 1 pour la classification de citation. La matrice de corrélation de celle-ci avec  $P$  s'étend de -0.17 à 0.54 et a une moyenne de valeur absolue de 0.08, ce qui est significatif par rapport à des classifications aléatoire puisque un bootstrap à  $b = 100$  répétitions avec les matrices mélangées donne un minimum à  $-0.08 \pm 0.012$ , un maximum à  $0.11 \pm 0.02$  et une moyenne absolue à  $0.03 \pm 0.002$ . Cela montre que les classifications sont complémentaires et que cette complémentarité est significative statistiquement par rapport à des classifications aléatoires. L'adéquation de la classification sémantique par rapport au réseau de citation peut également être quantifiée par la modularité multi-classes [NICOSIA et al., 2009] (voir C.5 pour une définition mathématique), qui traduit la probabilité qu'un lien soit dû à

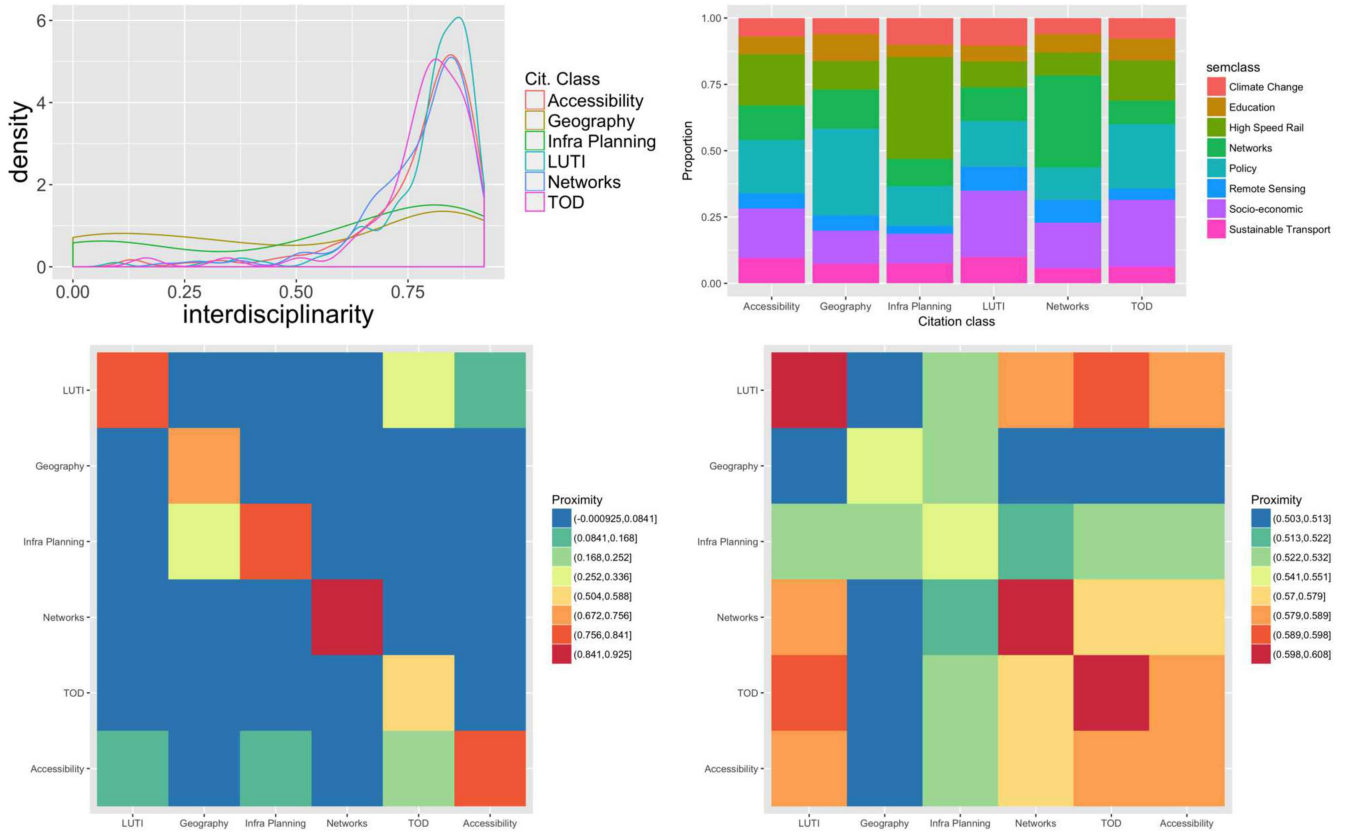


FIGURE 9 : **Motifs d'interdisciplinarité.** (*Haut Gauche*) Distribution statistique des  $I_i$  par classes de citations, en d'autre termes répartition des niveaux d'interdisciplinarité au sein des classes de citation ; (*Haut Droite*) Composition sémantiques des classes de citation : pour chaque classe de citation (en abscisse), la proportion de chaque classe sémantique (en couleur) est donnée ; (*Bas Gauche*) Matrice de proximité de citation  $c_{kk'}$  entre classes de citations ; (*Bas Droite*) Matrice de proximité sémantique  $s_{kk'}$  entre classes de citations.



la classification étudiée, en prenant en compte l'appartenance simultanée à de multiples classes. Ainsi, la modularité multi-classes des probabilités sémantiques pour le réseau de citation est de 0.10, ce qui d'une part est significativement signe d'adéquation, un bootstrap toujours à  $b = 100$  donnant une valeur de  $0.073 \pm 0.003$ , qui reste limitée vu la valeur maximale fixée par les probabilités de citations dans leur propre réseau qui donnent une valeur de 0.81, ce qui confirme d'autre part la complémentarité des classifications.

Nous avons ainsi dressé dans cette section un aperçu des disciplines en relation avec notre sujet, ainsi que leur relations. Il s'agira dans la section suivante de comprendre avec plus de détail leur "contenu", c'est-à-dire les moyens mobilisés pour résoudre les problèmes rencontrés.

### *Discussion*

Donnons brièvement des directions d'extension de l'analyse que nous venons de mener ainsi que des implications pour le positionnement épistémologique de notre travail.

#### *Vers une modélisation des thèmes et une extraction automatique du contexte*

Une direction possible pour renforcer cette analyse en épistémologie quantitative serait de travailler sur les textes complets des références contenant des efforts de modélisation des interactions entre réseaux et territoires, avec le but d'extraire automatiquement les thématiques des articles. Des méthodes plus adaptées pour les textes complets que celle utilisée ici incluent par exemple l'Allocation Latente de Dirichlet [BLEI, NG et JORDAN, 2003]. L'idée serait de procéder à une sorte de modélographie automatique, étendant la méthodologie de modélographie développée par [SCHMITT et PUMAIN, 2013], pour extraire des caractéristiques telle les ontologies, l'architecture ou la structure des modèles, les échelles ou même des valeurs typiques des paramètres. Il n'est pas clair dans quelle mesure la structure des modèles peut être extraite de leur description dans un article, et cela dépend sûrement de la discipline considérée. Par exemple dans un champ relativement cadré comme la planification des transports, l'utilisation d'une ontologie pré-définie (dans le sens d'un dictionnaire) et d'une grammaire floue pourrait être efficace vu les conventions assez strictes dans la discipline. En géographie théorique et quantitative, au-delà de la barrière de la diversité des formalisations possibles pour une même ontologie, l'organisation de l'information est sûrement plus délicate à appréhender par de l'apprentissage non-supervisé à cause de la nature plus littéraire de la discipline : les synonymes et les figures de style sont généralement la norme pour l'écriture d'un bon niveau en sciences humaines, rendant plus floue une possible structure générique de la description des connaissances.



*Réflexivité*

La méthodologie que nous avons développée ici est efficace pour offrir des instruments de réflexivité, c'est-à-dire qu'elle peut être utilisée pour étudier notre approche elle-même. Une de ses applications, hors de celle à la revue scientifique *Cybergeogéographie* dans la perspective de Science Ouverte (voir Annexe B.6), sera à notre propre corpus de références, dans le but de révéler des possibles directions de recherche ou problématiques exotiques. Il est éventuellement possible de le faire de manière dynamique, grâce à l'historique de git qui permet de récupérer n'importe quelle version de la bibliographie à une date donnée sur les trois ans écoulés. Il s'agira aussi de comprendre nos motifs de production de connaissance afin de contribuer à 8.3. Le développement détaillé est fait en Annexe F.

★ ★

★

Cette section nous a ainsi permis de dresser un paysage des disciplines en relation avec notre problématique, et des relations entre ces disciplines, en termes de citations mais aussi de niveau d'interdisciplinarité.

La section suivante se positionnera dans une démarche voisine, mais avec un but plus marqué d'exhaustivité en termes de modélisation des interactions : nous procéderons ainsi à une revue systématique et à une modélographie, afin de renforcer la typologie des modèles obtenue en section 2.1.

★ ★

★

## 2.3 REVUE SYSTÉMATIQUE ET MODÉLOGRAPHIE

Tandis que les études menées précédemment proposaient de construire un horizon global de l'organisation des disciplines s'intéressant à notre question, nous proposons à présent une étude plus ciblée des caractéristiques de modèles existants. Nous proposons pour cela dans un premier temps une revue systématique, c'est-à-dire la construction d'un corpus plus précis répondant à certaines contraintes, suivie d'une méta-analyse, c'est-à-dire une tentative d'explication de certaines caractéristiques des modèles par des modèles statistiques.

### 2.3.1 *Revue systématique*

Les revues systématiques classiques ont majoritairement lieu dans des domaines où une recherche très ciblée, même par titre d'article, fournira un certain nombre d'études étudiant quasiment la même question : typiquement en évaluation thérapeutique, où des études standardisées d'une même molécule varient uniquement par taille des effectifs et modalités statistiques (groupe de contrôle, placebo, niveau d'aveugle). Dans ce cas la construction du corpus est d'une part aisée par l'existence de bases spécialisées permettant des recherches très ciblées, et d'autre part par la possibilité de procéder à des analyses statistiques supplémentaires pour croiser les différentes études (par exemple méta-analyse par réseau, voir [RUCKER, 2012]). Dans notre cas, l'exercice est bien plus aléatoire pour les raisons exposées dans les deux sections précédentes : les objets sont hybrides, les problématiques diverses, et les disciplines variées. Les différents points soulevés par la suite auront souvent autant de valeur thématique que de valeur méthodologique, suggérant des points cruciaux lors de la réalisation d'une telle revue systématique hybride.

Nous proposons une méthodologie hybride couplant les deux méthodologies développées précédemment avec une procédure plus classique de revue systématique. Nous souhaitons à la fois une représentativité de l'ensemble des disciplines que l'on a découvertes, mais aussi un bruit limité dans les références prises en compte pour la modélographie. Pour cela, nous combinons le corpus obtenu précédemment et un corpus constitué par requêtes de mots-clés, de manière similaire à [TAHAMTAN et BORNMAN, 2018]. Le protocole est donc le suivant :

1. Partant du corpus de citation isolé en 2.2.3, nous isolons un nombre de mots-clés pertinents, en sélectionnant les 5% de liens ayant le plus fort poids (seuil arbitraire), puis parmi les noeuds correspondants ceux ayant un degré supérieur au quantile à 0.8 de leur classe sémantique respective. Le premier filtrage permet de se concentrer sur le "cœur" des disciplines observées, et le second de ne pas biaiser par la taille sans perdre la struc-

ture globale, les classes étant relativement équilibrées. Un examen manuel permet de supprimer les mots-clés clairement non-pertinents (télédetection, tourisme, réseaux sociaux, ...), ce qui conduit à un corpus de  $K = 115$  mots-clés ( $K$  est endogène ici).

2. Pour chaque mot-clé, nous effectuons automatiquement une requête au catalogue (scholar) en y ajoutant `model*`, d'un nombre fixé  $n = 20$  de références. L'ajout du terme est nécessaire pour obtenir des références pertinentes, après test sur des échantillons.
3. Le corpus potentiel composé des références obtenues, ainsi que des références composant le réseaux de citation, est revu manuellement (passage en revue des titres) pour assurer une pertinence au regard de l'état de l'art de 2.1, fournissant le corpus préliminaire de taille  $N_p = 297$ .
4. Ce corpus est alors inspecté pour les résumés et textes complets si nécessaire. On sélectionne les articles mettant en place une démarche de modélisation, hors modèles conceptuels. Les références sont classifiées et caractérisées selon des critères décrits ci-dessous. Nous obtenons alors un corpus final de taille  $N_f = 145$ , sur lequel des analyses quantitatives sont possibles.

La méthode est résumée en Fig. 10, avec les valeurs des paramètres et la taille des corpus successifs. Cet exercice permet tout d'abord un certain nombre de points méthodologiques, dont la connaissance pourra être un atout pour mener des revues systématiques hybrides similaires :

- Les biais de catalogue semblent inévitables. Nous reposons sur l'hypothèse que l'utilisation de Scholar permet un échantillonnage uniforme au regard des erreurs ou biais de catalogage. Le développement futur d'outils ouverts de catalogage et de cartographie, permettant un effort contributif pour une connaissance plus précise de domaines étendus et de leurs interfaces, sera un enjeu crucial de la fiabilité de ce genre de méthodes (voir B.6).
- La disponibilité des textes complets est particulièrement un problème pour une revue si large, vu la multiplicité des éditeurs. L'existence de moyens d'émancipation de la science ouverte comme Sci-hub<sup>34</sup> permet d'effectivement accéder à l'ensemble des textes. En écho au débat sur le bras de fer récent avec les éditeurs concernant l'exclusivité de la fouille de textes complets, il paraît de plus en plus évident qu'une science ouverte réflexive est totalement antagoniste au modèle actuel de l'édition. Nous espérons également une évolution rapide des pratiques sur ce point.

<sup>34</sup> <http://sci-hub.cc/>

- Les revues, et en fait les éditeurs, semblent influencer différemment le référencement, augmentant potentiellement le biais de requête. La littérature grise ainsi que les preprints sont pris en compte différemment selon les champs.
- Le passage en revue manuel des grands corpora permet de pas louper des “poids lourds” qui auraient pu être omis en amont [LISSACK, 2013]. La question de la mesure dans laquelle on peut s’attendre d’être au courant de la manière la plus exhaustive des découvertes récentes liées au sujet étudié évolue très probablement vu l’augmentation de la quantité totale de littérature produite et la fragmentation des domaines pour certains toujours plus pointus [BASTIAN, GLASZIOU et CHALMERS, 2010]. Rejoignant les points précédents, on peut supposer que des outils d’aide à l’analyse systématique permettront de garder cet objectif raisonnable.
- Les résultats de la revue automatique sont sensiblement différents des domaines dessinés dans la revue classique : certaines associations conceptuelles, notamment l’inclusion des modèles de croissance de réseaux, ne sont pas naturelles et existent peu dans le paysage scientifique comme nous l’avons montré précédemment.

D’autre part, l’opération de construction du corpus permet déjà de tirer des observations thématiques intéressantes en elles-mêmes.

- Les articles sélectionnés supposent une clarification de ce qui est entendu par “modèle”. Nous donnons en 8.3 une définition très large s’appliquant à l’ensemble des perspectives scientifiques. Notre sélection ici ne retient pas les modèles conceptuels par exemple, notre critère de choix étant que le modèle doit inclure un aspect numérique ou de simulation.
- Un certain nombre de références consistent en des revues, ce qui revient à un groupe de modèles ayant des caractéristiques similaires. Nous pourrions compliquer la méthode en retranscrivant chaque revue ou meta-analyse, ou en pondérant par le nombre d’article correspondant les enregistrements des caractéristiques correspondants. Nous faisons le choix d’ignorer ces revues, ce qui reste cohérent de manière thématique en restant dans l’hypothèse d’échantillonnage uniforme.
- Une première clarification du cadre thématique est opérée, puisque nous ne sélectionnons pas les études liées uniquement au trafic et à la mobilité (ce choix étant aussi lié aux résultats obtenus en ??), à l’urban design pur, aux modèles de flux piétons, au fret, à l’écologie, aux aspects techniques du transport, pour donner

quelques exemples, même si ces sujets peuvent dans une vue extrême être considérés comme liés aux interactions entre réseaux et territoires.

- De la même façon, des domaines annexes comme le tourisme, les aspects sociaux de l'accès aux transports, l'anthropologie, n'ont pas été pris en compte.
- On observe une forte fréquence des études liées au Trains à Grande Vitesse (HSR), rappelant la non-dissociabilité des aspects politiques de la planification et des directions de recherche en transports.

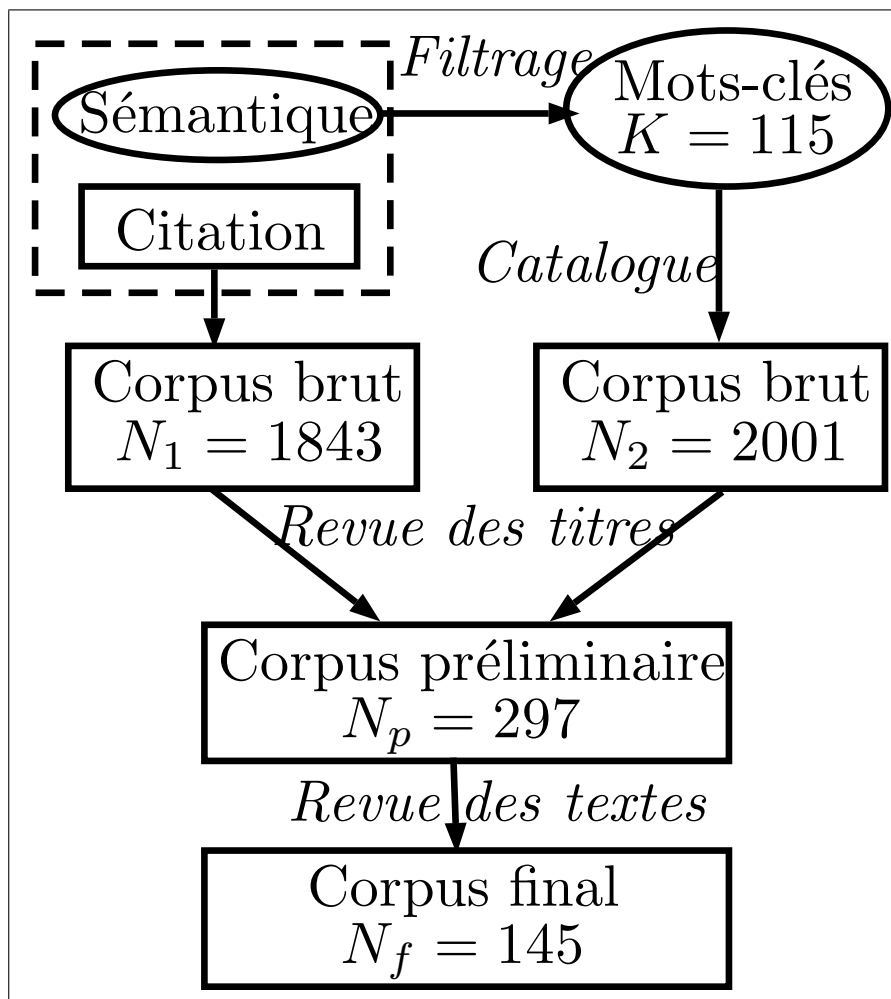


FIGURE 10 : **Méthodologie de la revue systématique.** Les rectangles désignent des corpus de références, les ellipses des corpus de mot-clés, et les pointillés les corpus initiaux. A chaque étape est donnée la taille du corpus.

### 2.3.2 Modélographie

Nous passons à présent à une analyse mixte basée sur ce corpus, inspirée par les résultats des sections précédentes précédents notamment pour la classification. Elle a pour but d'extraire et de décomposer précisément les ontologies, échelles et processus, puis d'étudier des liens possibles entre ces caractéristiques des modèles et le contexte dans lequel ils ont été introduits. Il s'agit ainsi de la méta-analyse en quelque sorte, que nous désignerons ici par modélographie. Pour ne pas froisser les puristes, il ne s'agit en effet pas d'une méta-analyse à proprement parler car nous ne combinons pas des analyses proches pour extrapoler des résultats potentiels d'échantillons plus grands. Notre démarche est proche de celle de [COTTINEAU, 2017] qui rassemble les références ayant étudié quantitativement la loi de Zipf pour les villes, puis lie les caractéristiques des études aux méthodes utilisées et hypothèses formulées.

La première partie consiste en l'extraction des caractéristiques des modèles. Automatiser ce travail constituerait un projet de recherche en lui-même, comme nous développons en discussion ci-dessous, mais nous sommes convaincu de la pertinence d'affiner de telles techniques (voir 8.3.3) dans le cadre d'un développement de disciplines intégrées. Le temps étant autant l'ennemi que l'allié de la recherche, nous nous concentrons ici sur une extraction manuelle qui se voudra plus fine qu'une tentative peu convaincante de fouille de données. Nous extrayons des modèles les caractéristiques suivantes :

- quelle est la force du couplage<sup>35</sup> entre les ontologies territoriales et celles du réseau, autrement dit s'agit-il d'un modèle de co-évolution. Nous classerons pour cela en catégories suivant la représentation de la figure 11 : {territory ; network ; weak ; coevolution}, qui résulte de l'analyse de la littérature en 2.1 ;
- échelle de temps maximale ;
- échelle d'espace maximale ;
- hypothèses d'équilibre ;
- domaine "a priori", déterminé par l'origine des auteurs et domaine de la revue ;
- méthodologie utilisée (modèles statistiques, système d'équations, multi-agent, automate cellulaire, recherche opérationnelle, simulation etc.) ;

<sup>35</sup> Il n'existe à notre connaissance pas d'approche générique du couplage des modèles n'étant pas liée à un formalisme particulier. Nous prendrons l'approche donnée en introduction, en distinguant ici un couplage faible comme un couplage séquentiel (sorties du premier modèle deviennent entrées du second) d'un couplage fort dynamique où l'évolution est interdépendante à chaque instant (soit par une détermination réciproque soit par une ontologie commune).

- cas d'étude (ville, métropole, région ou pays) s'il y a lieu.

Nous collectons également de manière indicative, mais sans objectif d'objectivité ni d'exhaustivité, le "sujet" de l'étude (c'est-à-dire la question thématique dominante) ainsi que les "processus" inclus dans le modèle. Une extraction exacte des processus reste hypothétique, d'une part conditionnée à une définition rigoureuse et prenant en compte différents niveaux d'abstraction, de complexité, ou d'échelle, d'autre part dépendant de moyens techniques hors de portée de cette étude modeste. Nous commenterons ceux-ci de manière indicative sans les inclure dans les études systématiques.

Nous confondons également échelle, portée et dans un sens résolution pour ne pas rendre plus confus l'extraction. Même s'il serait pertinent de différencier lorsque un élément n'a pas lieu d'être pour un modèle (NA) de lorsque celui-ci est mal défini par son auteur, cette tâche apparaît sujette à subjectivité et nous fusionnons les deux modalités. Nous ajoutons aux caractéristiques ci-dessus les variables suivantes :

- domaine de citation (le cas échéant, c'est-à-dire pour les références initialement présentes dans le réseau de citation, i.e. 55% des références);
- domaine sémantique, défini par le domaine pour lequel le document a la plus grande probabilité;
- indice d'interdisciplinarité.

Les domaines sémantiques et la mesure d'interdisciplinarité ont été recalculés pour ce corpus par collecte des mots-clés, puis extraction selon la méthode décrite en 2.2, avec  $K_W = 1000$ ,  $\theta_w = 15$  et  $k_{max} = 500$ . On obtient des communautés plus ciblées et plutôt représentatives de la thématique et des méthodes : Transit-oriented development (tod), Hedonic models (hedonic), Planification des infrastructures (infra planning), High-speed rail (hsr) , Réseaux (networks), Réseaux complexes (complex networks), Bus rapid transit (brt).

Un "bon choix" de caractéristiques pour classer les modèles est un peu le problème du choix des *features* en apprentissage statistique : si on est en supervisé, c'est-à-dire qu'on veut obtenir une bonne prédiction de classe fixée a priori (ou une bonne modularité de la classification obtenue par rapport à la classification fixée), on pourra sélectionner les caractéristiques optimisant cette prédiction. On discriminera ainsi les modèles que l'on connaît et que l'on juge différents. Si l'on veut extraire une structure endogène sans a priori (classification non supervisée), la question est différente. Nous testerons pour cela en second temps une technique de regression qui permet d'éviter l'overfitting et faire de la selection de caractéristiques (forêts aléatoires).



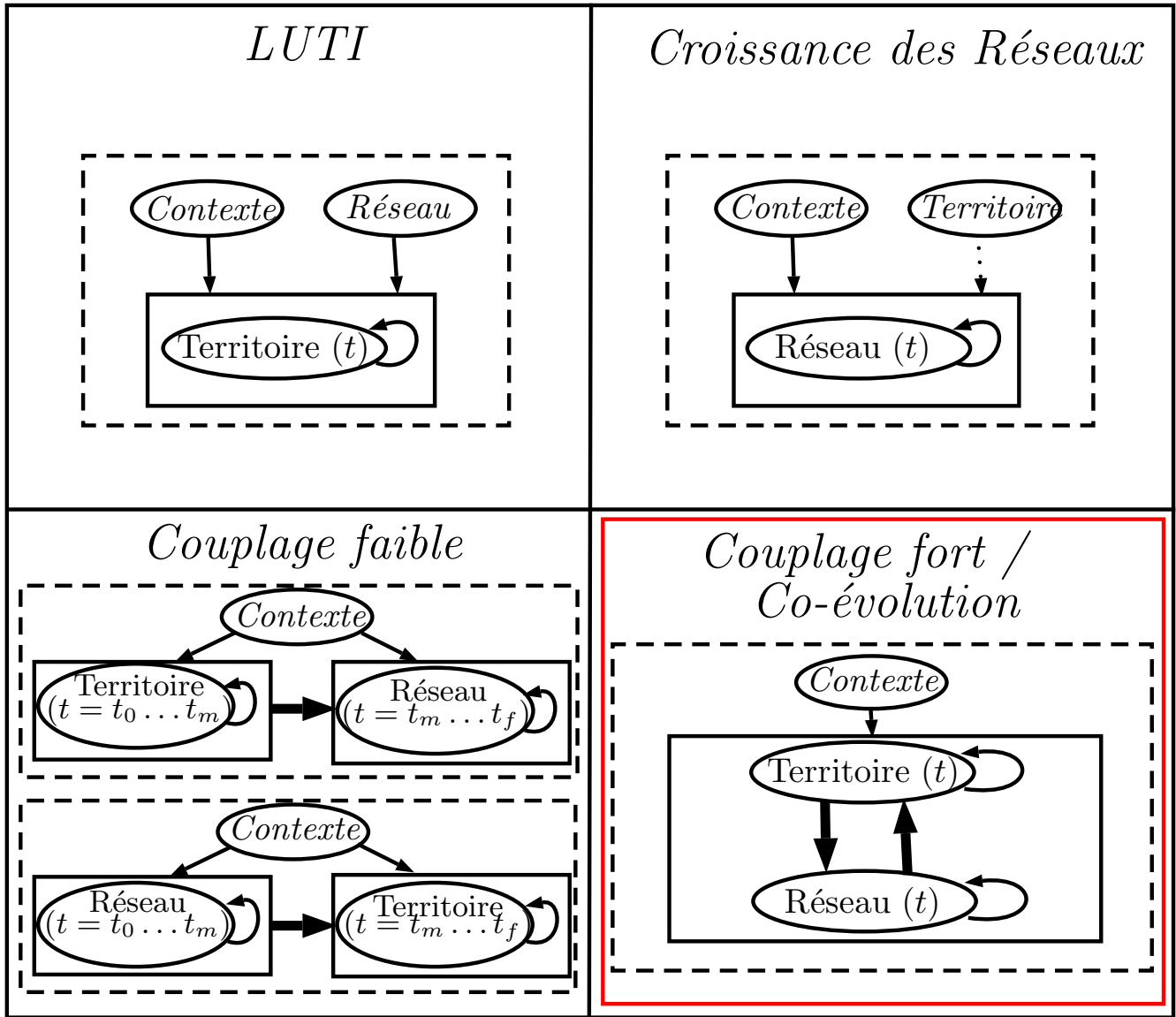


FIGURE 11 : Représentation schématique de la distinction entre différents types de modèles couplant territoires et réseaux. Cette typologie reprend celle de la Table 4, en distinguant les approches où territoire ou réseau sont donnés comme contexte (Luti et croissance du réseau) d'un couplage séquentiel entre un modèle pour chaque. Les ontologies sont représentés par des ovales, les sous-modèles par les boîtes pleines, les modèles par les boîtes pointillées, les couplages par les flèches. Nous surlignons en rouge l'approche qui sera l'objectif final de notre travail.

#### Processus et cas d'étude

Concernant l'existence d'un cas d'étude et sa localisation, 26% des études n'en présentent pas, correspondant à un modèle abstrait ou modèle jouet (la quasi totalité des études en physique tombant dans ce cas). Ensuite, elles sont réparties à travers le monde, avec toutefois une surreprésentation des Pays-bas avec 6.9%. Les processus in-

clus sont trop variés (en fait autant que les ontologies des disciplines concernées) pour faire l'objet d'une typologie, mais nous noterons la domination de la notion d'accessibilité (65% des études), puis des processus très variés allant de processus de marché immobilier pour les études hédoniques, aux relocalisations d'actifs et d'emplois pour les Luti, ou aux investissements d'infrastructure de réseau. Nous observons des processus abstraits géométriques de croissance de réseau, correspondant aux travaux des physiciens. La maintenance du réseau apparaît dans une étude, ainsi que l'histoire politique. Les processus abstraits d'agglomération et dispersion sont aussi le cœur de quelques études. Les interactions entre villes sont minoritaires, les approches de type systèmes de villes étant noyées dans les études d'accessibilité. Les questions de gouvernance et de régulation ressortent aussi, plutôt dans le cas de planification d'infrastructure et de modèle d'évaluation de démarches TOD, mais sont aussi minoritaires. On retiendra que chaque domaine puis chaque étude introduit ses propres processus quasi-spécifiques à chaque cas.

#### *Caractéristiques du corpus*

Les domaines "a priori" (i.e. jugés, ou plutôt préjugés sur la revue ou l'appartenance des auteurs), sont relativement équilibrés pour les disciplines majoritaires déjà identifiées : 17.9% Transportation, 20.0% Planning, 30.3% Economics, 19.3% Geography, 8.3% physics, le reste minoritaire se répartissant entre environnement, informatique, ingénierie et biologie. Concernant les poids des domaines sémantiques significatifs, le TOD domine avec 27.6% des documents, suivi par les réseaux (20.7%), les modèles hédoniques (11.0%), la planification des infrastructures (5.5%) et le HSR (2.8%). Les tables de contingences montrent que le Planning ne fait quasiment que du TOD, la physique uniquement des réseaux, la géographie se répartit équitablement entre réseaux et TOD (le second correspondant aux articles typés "aménagement", qui ont été classés en géographie car dans des revues de géographie) ainsi qu'une plus faible part en HSR, enfin l'économie est la plus variée entre hédonique, planning, réseaux et TOD. Cette interdisciplinarité n'apparaît cependant que pour les classes extraites pour la probabilité majoritaire, puisque les indices d'interdisciplinarité moyens par discipline ont des valeurs équivalentes (de 0.62 à 0.65), hormis la physique significativement plus basse à 0.56 ce qui confirme son statut de "nouveau venu" ayant une profondeur thématique plus faible.

#### *Modèles étudiés*

Il est intéressant pour notre question de répondre à la question "qui fait quoi?", c'est-à-dire quelles types de modèles sont mobilisés par les différentes disciplines. Nous donnons en Table 8 la table de contin-

TABLE 8 : **Types de modèles étudiés selon les différentes classifications.** Tables de contingence de la variable discrète donnant le type de modèle (réseau, territoire ou couplage fort), pour la classification a priori, la classification sémantique et la classification de citation.

Discipline	economics	geography	physics	planning	transportation
network	5	3	12	1	4
strong	4	3	0	0	2
territory	35	22	0	28	20
Semantic	hedonic	hsr	infra ning	plan- networks	tod
network	1	0	0	14	2
strong	0	0	0	5	1
territory	15	4	8	11	37

Citation	accessibility	geography	infra ning	Plan- LUTI	networks	TOD
network	0	0	0	0	24	0
strong	0	0	0	2	5	0
territory	13	1	6	18	2	3

gence du type de modèle en fonction des disciplines a priori, de la classe de citation et de la classe sémantique. On constate les approches fortement couplées, les plus proches de ce qu'on considère comme des modèles de co-évolution, sont majoritairement contenues dans le vocabulaire des réseaux, ce qui est confirmé par leur positionnement en terme de citation, mais que les disciplines concernées sont variées. La majorité des études s'intéresse au territoire uniquement, le déséquilibre le plus fort étant pour les études sémantiquement liées au TOD et à l'hédonique. La physique est encore limitée en s'intéressant exclusivement aux réseaux.

#### *Échelles étudiées*

Pour répondre ensuite à la question du comment, on peut regarder les échelles de temps et d'espace typiques des modèles. La planification et les transports se concentrent à des petites échelles spatiales, métropolitain ou local, l'économie également avec une forte représentation du local via les études hédoniques, et une étendue un peu plus grande avec l'existence d'études au niveau régional et quelques une du pays (études de panel généralement). Encore une fois, la physique se retrouve limitée avec l'ensemble de ses contributions à une échelle fixe, métropolitaine (pas forcément claire ni bien spécifiée dans les articles d'ailleurs puisqu'il s'agit de modèles jouets dont les contours thématiques peuvent être très flous). La géographie est relativement

bien équilibrée, de l'échelle métropolitaine à l'échelle continentale. Le schéma pour les échelles de temps est globalement similaire. Les méthodes utilisées sont fortement corrélées à la discipline : un test du  $\chi^2$  donne une statistique de 169, très significatif avec  $p = 0.04$ . De même, l'échelle d'espace l'est mais de manière moindre ( $\chi^2 = 50$ ,  $p = 0.08$ ).

### *Régressions classiques*

Nous étudions à présent l'influence de divers facteurs sur les caractéristiques des modèles par des régressions linéaires simples. Dans une démarche de multi-modélisation, nous proposons de tester l'ensemble des modèles possibles pour expliquer chacune des variables à partir des autres. Le nombre d'observations pour lesquelles toutes les variables sont renseignées est très faible, il s'agit de prendre en compte le nombre d'observations utilisées pour ajuster chaque modèle. D'autre part, les performances du modèle peuvent être caractérisées par des objectifs complémentaires. Suivant [IGEL, 2005], nous appliquons une optimisation multi-objectif, pour maximiser simultanément la variance expliquée ( $R^2$  ajusté dans notre cas) et l'information capturée (Critère d'information d'Akaike corrigé AICc<sup>36</sup>). Celle-ci est effectuée conditionnellement au fait d'avoir le nombre d'observations  $N > 50$  (seuil fixé au regard de la distribution de  $N$  sur l'ensemble des modèles). La procédure d'optimisation est détaillée en Annexe A.2 pour chaque variable. L'échelle de temps et l'interdisciplinarité présentent des compromis difficiles à départager, et nous ajustons les deux candidats. Les autres variables présentent des solutions dominantes et nous n'ajustons qu'un seul modèle.

Les résultats complets des régressions sont donnés en Table 9. Les échelles temporelle et d'espace, ainsi que l'année, sont les variables les mieux expliquées au sens de la variance. L'échelle de temps est influencée très significativement par le type de modèle : territoire qui diminue celle-ci, ou couplage fort qui l'augmente. Le fait d'être en physique influe également significativement, et élargit la portée temporelle des modèles. Au contraire, les approches d'ingénierie (souvent design optimal d'un réseau de transport) correspondent à une courte durée.

Pour l'échelle d'espace, le fait d'être en géographie a une forte influence sur la portée spatiale des modèles : en effet, les études régionales et à l'échelle du système de villes sont bien l'apanage de la géographie. L'appartenance au domaine du transport augmente aussi faiblement la portée spatiale (voir significativité dans les régressions complètes en Annexe A.2). Aucune autre variable n'a une influence significative.

<sup>36</sup> L'AIC est une mesure du gain d'information entre deux modèles, et permet d'éviter l'ajustement abusif par un nombre trop grand de paramètres. L'AICc est une version prenant en compte la taille de l'échantillon, la mesure variant significativement pour les petits échantillons.

Le niveau d'interdisciplinarité est bien expliqué par l'année, qui l'influence de manière négative, ce qui confirme une augmentation des spécialisations scientifiques dans le temps. Les études économétriques des modèles hédoniques apparaissent très spécialisées. Enfin, l'année de publication est expliquée significativement et positivement par le type territoire et par le fait d'être en transports, ce qui signifierait une recrudescence récente d'un profil particulier d'études. Un examen du corpus suggère qu'il s'agit des études sur la grande vitesse, apparaissant comme une mode scientifique récente.

#### *Régressions par Forêts Aléatoires*

Nous concluons cette étude par des régressions et classification par forêts aléatoires, qui sont une méthode très flexible permettant de dégager une structure d'un jeu de données [LIAW et WIENER, 2002]. Pour compléter les analyses précédentes, nous proposons de l'utiliser pour déterminer les importances relatives des variables pour différents aspects. Nous utilisons à chaque fois des forêts de taille 100000, une taille de noeud de 1 et un nombre de variable échantillonnée en  $\sqrt{p}$  pour la classification et  $p/3$  pour la régression lorsque  $p$  est le nombre total de variables. Pour classifier le type de modèle, nous comparons les effets de la discipline, de la classe sémantique et de la classe de citation. Cette dernière est la plus importante avec une mesure relative de 45%, tandis que la discipline compte pour 31% et le sémantique pour 23%. Ainsi, le cloisonnement disciplinaire se retrouve, tandis que le sémantique et donc en partie les ontologies, est le plus ouvert. Cela nous encourage dans notre démarche de sortir de ce cloisonnement. Lorsqu'on applique une régression de forêt sur l'interdisciplinarité, toujours avec ces trois variables, on constate qu'elles expliquent 7.6% de la variance totale, ce qui est relativement faible, témoignant d'une disparité de sémantique sur l'ensemble du corpus indépendamment des différentes classifications. Dans ce cas, la variable la plus importante est la discipline (39%) suivie par le sémantique (31%) et la citation (29%), ce qui confirme que le journal visé conditionne fortement le comportement de langage employé. Cela nous alerte sur le danger de perte de richesse sémantique lorsqu'on s'adresse à un public particulier. Ainsi, nous avons pu dégager certaines structures et régularités des modèles nous concernant, qui seront riches d'enseignements lors de la construction de nos modèles.

#### 2.3.3 *Discussion*

##### *Développements*

Un développement possible pourrait consister en la mise en place d'une approche automatique à cette méta-analyse, du point de vue de la modélisation modulaire, combiné avec une classification du but

TABLE 9 : **Explication des caractéristiques des modèles.** Résultats de l'estimation par moindres carrés (OLS) des modèles linéaires sélectionnés, pour chacune des variables à expliquer : échelle temporelle (TEMPSCALE), échelle spatiale (SPATSCALE), indice d'interdisciplinarité (INTERDISC), année de publication (YEAR).

	<i>Variable expliquée :</i>					
	TEMPSCALE		SPATSCALE	INTERDISC		YEAR
	(1)	(2)	(3)	(4)	(5)	(6)
YEAR	0.674			−0.004*	−0.002*	
TYPEstrong		100.271***			−0.026	
TYPEterritory	−38.933***	−14.988			0.044	10.898***
TEMPSCALE			−5.179	−0.0003		0.035
FMETHODeq						−6.224
FMETHODmap						4.747
FMETHODro						6.128
FMETHODsem						1.009
FMETHODsim						5.153
FMETHODstat						−0.357
DISCIPLINEengineering	−52.107*	−9.609	−154.461	0.144		13.486
DISCIPLINEenvironment	17.110	17.886	−5.878	0.092		−3.668
DISCIPLINEgeography	3.640	9.126	1,445.457***	0.036		1.121
DISCIPLINEphysics	46.879*	77.897***	292.559	−0.103		3.392
DISCIPLINEplanning	1.304	4.553	−143.554	−0.047		−2.850
DISCIPLINEtransportation	−14.718	8.753	568.329	0.062		5.503*
INTERDISC	2.357					−12.876
SEMCOMcomplex networks					−0.217	
SEMCOMhedonic				−0.179	−0.184*	−5.769
SEMCOMhsr				−0.100	−0.122	6.135
SEMCOMinfra planning				−0.032	−0.096	−4.123
SEMCOMnetworks				−0.038	−0.107	4.711
SEMCOMtod				−0.105	−0.152	−1.653
Constant	−1,305.126	22.103*	235.357	8.962**	5.531**	2,004.945***
Observations	64	94	94	64	98	64
R <sup>2</sup>	0.385	0.393	0.100	0.314	0.155	0.510
R <sup>2</sup> ajusté	0.282	0.336	0.027	0.136	0.068	0.281

Note :

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

et de l'échelle. La modélisation modulaire consiste en l'intégration de processus hétérogènes et d'implémentation de ces processus dans le but d'extraire les mécanismes donnant la meilleure proximité à des faits stylisés empiriques ou à des données [COTTINEAU, CHAPRON et REUILLON, 2015]. L'idée serait de pouvoir extraire automatiquement la structure modulaire des modèles existants, à partir des textes complets comme proposé en 2.2, afin de classer ces briques de manière endogène et identifier des couplages potentiels pour des nouveaux modèles.

#### *Leçons pour la modélisation*

Nous pouvons résumer les points principaux issus de cette méta-analyse qui joueront sur notre attitude et nos choix de modélisation. Tout d'abord, la présence interdisciplinaire des approches effectuant un couplage fort confirme notre besoin de faire des ponts et de coupler les approches, et confirme également rétrospectivement les conclusions de 2.2 sur les conséquences du cloisonnement des disciplines en terme de modèles formulés. Ensuite, l'importance du vocabulaire des réseaux dans une grande partie des modèles nous poussera à confirmer cet ancrage. La spécificité des approches TOD et d'accessibilité, assez proches des modèles LUTI, seront secondaires pour nous. La portée restreinte des travaux issus de la physique, confirmée par la majorité des critères étudiés, nous pousse à nous méfier de ces travaux et de l'absence de sens thématique aux modèles. La richesse des échelles temporelles et spatiales couvertes par les modèles géographiques et économiques nous confirme l'importance de varier celles-ci dans nos modèles, idéalement de parvenir à des modèles multi-échelles. Enfin, les importances relatives des variables de classification sur le type de modèle vont également dans le sens de ponts interdisciplinaires pour croiser les ontologies.

★ ★

★



TABLE 10 : **Synthèse des processus modélisés.** Ceux-ci sont classés par échelle, type de modèle et discipline.

	Réseaux → Territoires	Territoires → Réseaux	Réseaux ↔ Territoires
Micro	<b>Economie</b> : marché immobilier, relocalisation, marché de l'emploi	NA	<b>Informatique</b> : croissance spontanée
	<b>Planification</b> : régulations, développement		
Meso	<b>Economie</b> : marché immobilier, coût du transport, aménités	<b>Physique</b> : corrélations topologiques, hiérarchie, congestion, optimisation locale, maintenance du réseau	<b>Economie</b> : investissements, relocalisations, offre et demande, planification du réseau
	<b>Géographie</b> : usage du sol, centralité, étalement urbain, effets de réseau	<b>Transports</b> : investissements, niveau de gouvernance	<b>Géographie</b> : usage du sol, croissance du réseau, diffusion de population
	<b>Planification/transports</b> : accessibilité, usage du sol, relocalisation, marché immobilier	<b>Economie</b> : croissance du réseau, offre et demande	
Macro	<b>Géographie</b> : accessibilité, interaction entre villes, relocalisation, histoire politique	<b>Géographie</b> : interactions entre villes, investissements	<b>Economie</b> : offre et demande
	<b>Transports</b> : accessibilité, marché immobilier	<b>Transports</b> : planification de réseau	<b>Transports</b> : couverture du réseau
	<b>Economie</b> : croissance économique, marché, usage du sol, agglomération, dispersion, compétition	<b>Géographie</b> : interactions entre villes, rupture de potentiel	

## SYNTHÈSE DES PROCESSUS MODÉLISÉS

Nous proposons de synthétiser les processus pris en compte par les modèles parcourus lors de la modélographie, afin de procéder à un effort similaire à celui concluant l'approche thématique du chapitre 1. Nous ne pouvons ni avoir une vision exhaustive (comme déjà précisé lors de la description de la méthodologie de la modélographie) ni rendre compte avec grande précision de chaque modèle en détail, puisque quasiment chacun est unique dans son ontologie. L'exercice de synthèse permet ainsi de s'extraire de ces limites et prendre un certain recul, et avoir ainsi un aperçu sur les *processus modélisés*<sup>37</sup>.

37 En gardant en tête les choix de sélection, qui emmènent par exemple à ne pas avoir les processus de mobilité dans cette synthèse.

La table 10 propose cette synthèse à partir des 145 articles issus de la modélographie et pour lesquels une classification de type était possible, c'est-à-dire qu'il existait un modèle rentrant dans la typologie développée en 2.1. Être complètement exhaustif relèverait d'une opération de méta-modélisation interdisciplinaire qui est bien hors de la portée de notre travail<sup>38</sup>, et la liste donnée ici est indicative.

Nous retrouvons les correspondances entre disciplines, échelles et types de modèles obtenues dans la modélographie en 2.3. Nous retirons les enseignements principaux suivants, en écho au tableau de synthèse obtenu en fin du Chapitre 1 (Table 3) :

1. La dichotomie des ontologies et des processus pris en compte entre les échelles et entre les types est autant manifeste ici dans les modèles que dans les processus en eux-même<sup>39</sup>. Nous postulons qu'il existe bien des processus différents aux différentes échelles, et nous prendrons le parti d'étudier différentes échelles.
2. Le cloisonnement des disciplines démontré en 2.2 se retrouve qualitativement dans cette synthèse : il est évident qu'elles divergent originellement dans leurs différentes épistémologies fondatrices. Nous tacherons d'intégrer des paradigmes de différentes disciplines, tout en prenant en compte les limites imposées par les principes de modélisation que nous présenterons en 3.1 (par exemple, la parcimonie des modèles limite nécessairement l'intégration d'ontologies hétérogènes).
3. Un décalage important entre cette synthèse et celle des processus est la quasi absence ici de modèles intégrant des processus de gouvernance. Il s'agira d'une piste à explorer.
4. Au contraire, une très bonne correspondance s'établit entre les modèles géographiques des systèmes urbains et les positionnements théoriques de la théorie évolutive des villes. Cette adéquation, plus difficile à retrouver pour l'ensemble des autres approches revues, nous suggère également de suivre cette piste.

<sup>38</sup> Il s'agirait pour cela d'avoir des correspondances entre les ontologies, sans lesquelles on se retrouverait avec au moins autant de processus que de modèles, même au sein d'une discipline. Il n'en existe à notre connaissance pas entre deux disciplines seulement. Une piste pour une approche formelle est donnée en B.5.

<sup>39</sup> Puisqu'on a plus détaillé cette étude, elle paraît même plus forte aussi, une plus grande précision permettant alors de séparer des catégories abstraites.

## CONCLUSION DU CHAPITRE

Les processus que nous cherchons à modéliser étant multi-scalaires, hybrides et hétérogènes, les angles d’approches et questionnements possibles sont nécessairement extrêmement variés, complémentaires et riches. Il pourrait s’agir d’une caractéristique fondamentale des systèmes socio-techniques, que PUMAIN formule dans [PUMAIN, 2005] comme “une nouvelle mesure de complexité”, qui serait liée aux nombre de point de vue nécessaires pour appréhender un système à un niveau donné d’exhaustivité. Cette idée rejoint la position de *perspectivisme appliqué* que B.5 formalise et qui est implicitement présente dans l’investigation des relations entre économie et géographie développée en C.6. Ainsi, la modélisation des interactions entre réseaux et territoires peut être reliées à un ensemble très large de disciplines et d’approches revues en section 2.1.

Afin de mieux comprendre le paysage scientifique environnant, et quantifier les rôles ou poids relatifs de chacune, nous avons procédé à une série d’analyse en épistémologie quantitative en 2.2. Une première analyse préliminaire basée sur une revue systématique algorithmique suggère un certain cloisonnement des domaines. Cette conclusion est confirmée par l’analyse d’hyperréseau couplant réseau de citations et réseau sémantique, qui permet également de dessiner plus finement les contours disciplinaires, à la fois sur leur relations directes (citations) mais aussi leur proximité scientifique pour les termes et méthodes utilisées. On peut alors utiliser le corpus constitué et cette connaissance des domaines pour une revue systématique semi-automatique en 2.3, qui permet de constituer un corpus de travaux traitant directement du sujet, qui est ensuite inspecté intégralement, permettant de lier caractéristique des modèles au différents domaines. Nous avons alors à ce stade une idée assez précise de ce qui se fait, pourquoi et comment.

L’enjeu reste de déterminer les pertinences relatives de certaines approches ou ontologies, ce qui sera le but des deux chapitres de la deuxième partie. Nous concluons d’abord cette première partie par un chapitre de discussion 3, éclairant des points nécessaires à clarifier avant une entrée dans le vif du sujet.

★                      ★

★