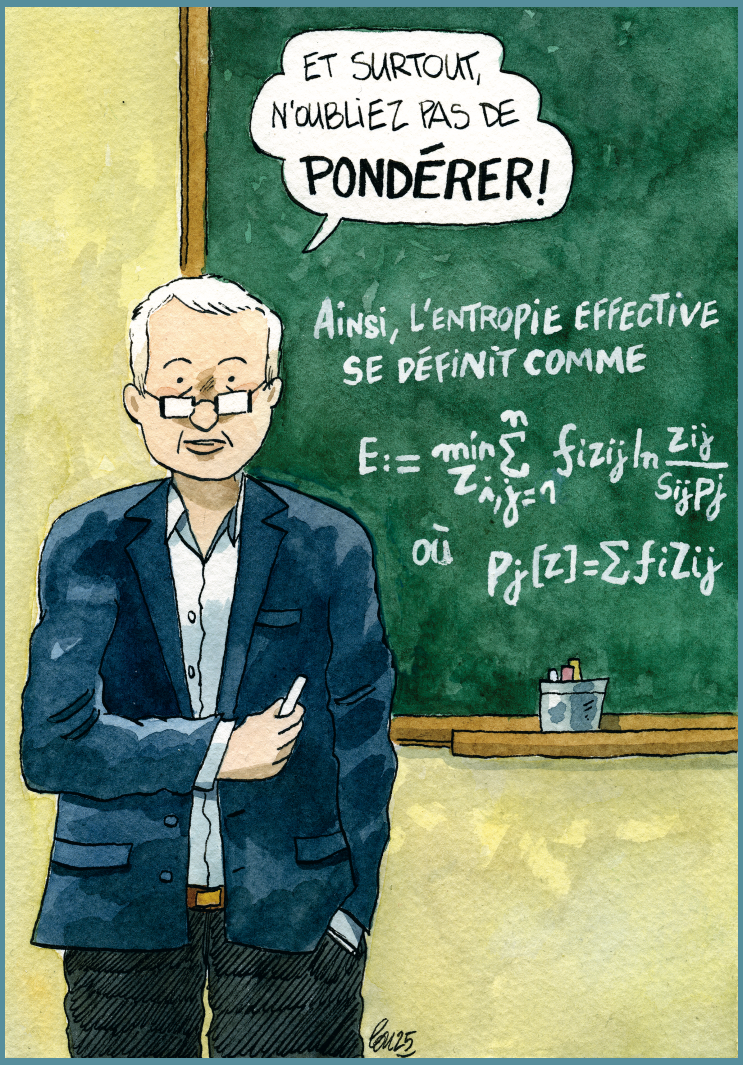
Espace, langage, réseaux, science
mélanges offerts en hommage à François Bavaud

Édités par Romain Loup et Aris Xanthos



Cahiers du CLSL, n°69, 2025

UNIL | Université de Lausanne

Espace, langage, réseaux, science :
mélanges offerts en hommage à François Bavaud

Édités par
Romain Loup et Aris Xanthos

Cahiers du CLSL, n°69, 2025

Illustration de couverture par Guillaume Long

Ont déjà paru dans cette série :

Discours sur les langues et rêves identitaires (2009, n° 26)
Langue et littératures pour l'enseignement du français en Suisse romande : problèmes et perspectives (2010, n°27)
Barrières linguistiques en contexte médical (2010, n° 28)
Russie, linguistique et philosophie (2011, n° 29)
Plurilinguismes et construction des savoirs (2011, n° 30)
Langue(s). Langage(s). Histoire(s). (2011, n° 31)
Identités en confrontation dans les médias (2012, n° 32)
Humboldt en Russie (2013, n° 33)
L'analyse des discours de communication publique (2013, n° 34)
L'édification linguistique en URSS : thèmes et mythes (2013, n° 35)
Mélanges offerts en hommage à Remi Jolivet (2013, n° 36)
Histoire de la linguistique générale et slave : sciences et traditions (2013, n° 37)
Ireland and its Contacts/L'Irlande et ses contacts (2013, n° 38)
La linguistique urbaine en Union Soviétique (2014, n° 39)
La linguistique soviétique à la recherche de nouveaux paradigmes (2014, n° 40)
Le niveau méso-interactionnel : lieu d'articulation entre langage et activité (2014, n° 41)
L'expertise dans les discours de la santé. Du cabinet médical aux arènes publiques, (2015, n°42)
L'école phonologique de Leningrad : histoire et modernités, (2015, n°43)
Le malentendu dans tous ses états, (2016, n°44)
Nouvelles technologies et standards méthodologiques en linguistique, (2016, n°45)
Aleksandr Potebnja, langage, pensée, (2016, n°46)
Rozalija Šor (1894-1939) et son environnement académique et culturel, (2016, n°47)
Perspectives on English in Switzerland, (2016, n°48)
Cinquante nuances du temps et de l'espace dans les théories linguistiques, (2016, n°49)
Le palimpseste gotique de Bologne. Etudes philologiques et linguistiques, (2016, n°50)
Les communautés suisses de Crimée et de la mer Noire : Langues et traditions, (2017, n°51)
Historiographie & épistémologie des sciences du langage : du passé vers le présent (2018, n°52)
Linguistique et philosophie du langage, (2018, n°53)
Investigating journalism practices (2018, n°54)
La communication digitale 1 (2018, n°55)
Mélanges offerts en hommage à Marianne Kilani-Schoch (2018, n°56)
Le Cours de linguistique générale : réception, diffusion, traduction (2018, n°57)
La médiation des savoirs sur le langage (2019, n°58)
Se mettre en scène en ligne. La communication digitale 2 (2019, n°59)
Hommage à Rudolph Wachter (2019, n°60)
Interlinguistique et espérantologie (2019, n°61)
Méthodes et modèles de l'apprentissage des langues anciennes, vivantes et construites, hier et aujourd'hui (2020, n°62)
Langues agglutinantes (2021, n°63)
Les communautés en ligne (2021, n°64)
Histoire des sciences du langage : approche épistémologique (2021, n°65)
La traduction et son processus didactique (2022, n°66)
De la théorie des Matrices et des Etymons à la submorphologie motivée (2024, n°67)
Language Pedagogy from Memphis to Tokyo (2024, n°68)

www.cahiers-clsl.ch

Espace, langage, réseaux, science :
mélanges offerts en hommage à François Bavaud

Édités par
Romain Loup & Aris Xanthos

Cahiers du CLSL, n°69, 2025

UNIL | Université de Lausanne

Les Cahiers du CLSL
(ISBN 978-2-940607-19-8)
Centre de Linguistique et des Sciences du Langage
Université de Lausanne (Suisse)
www.cahiers-clsl.ch

Centre de Linguistique et des Sciences du Langage
Quartier UNIL-Dorigny, Bâtiment Anthropole
CH-1015 Lausanne

Table des matières

Cahiers du CLSL, n° 69, 2025, pp. 7-17

Avant-propos

Aris Xanthos & Romain Loup
Université de Lausanne
{aris.xanthos,romain.loup}@unil.ch

L'édition d'un *Festschrift* est un exercice qui présente des points communs avec la pêche au chalut. On lance ses filets sans jamais savoir au juste ce que l'on y recueillera : poissons déjà repérés ou ramenés par des courants inattendus ; espèces familières ou venues de récifs lointains ; analyses approfondies, reprises de dialogues anciens ou échos de remous intellectuels que la personne honorée a contribué à faire naître. On sait en tout cas que la récolte, quelle qu'en soit la composition, sera le reflet d'un écosystème particulier – ici, celui dans lequel François Bavaud a évolué, enseigné, publié, influencé. Comme le chalut qui racle les fonds marins pour en rapporter la diversité vivante, ce volume ramène à la surface la richesse d'un environnement intellectuel, académique et institutionnel marqué par la présence de François.[1]

Espace, langage, réseaux, science : voici, nous semble-t-il, une énumération concise des grandes thématiques qui structurent cette collection d'articles et reflètent des aspects centraux, récurrents et interconnectés de l'univers intellectuel de François Bavaud. Espace et langage sont naturellement des objets fondamentaux dans ses deux facultés de rattachement à l'Université de Lausanne (UNIL) – Géosciences et Lettres – et dont il a maintes fois exploré les aspects formels et mathé-

1 L'édition d'un *Festschrift* est aussi un travail d'équipe. A ce titre, les éditeurs tiennent à remercier chaleureusement Laura Delaloye Saillen pour son aide et sa relecture précieuses, Guillaume Long pour le dessin de la couverture, le Centre de linguistique et des sciences du langage, la Section des sciences du langage et de l'information et l'Institut de géographie et durabilité (tous trois à l'Université de Lausanne) pour le financement de cet ouvrage, enfin l'ensemble des auteur.e.s pour leurs contributions et leur participation à la relecture.

matiques. En témoignent notamment ses travaux autour de l'autocorrélation spatiale (cf. Bavaud, 2005, 2008b, 2013, 2024; Ceré & Bavaud, 2017; Loup & Bavaud, 2025), de l'analyse de flux (cf. Bavaud, 2008a; Bavaud et al., 2018; Guex & Bavaud, 2015; Guex et al., 2023) et de l'application aux données textuelles ou linguistiques de méthodes factorielles, markoviennes et basées sur la théorie de l'Information (cf. Bavaud, 2018; Bavaud et al., 2006; Bavaud & Xanthos, 2002, 2005; Cocco et al., 2011; Egloff & Bavaud, 2018). Considérations spatiales et linguistiques se rejoignent d'ailleurs dans certaines publications (Bavaud et al., 2012, 2015; Ceré et al., 2018). Comme on peut s'y attendre, les deux domaines sont bien représentés dans ce volume, au travers des contributions de Bivand, Jeanson, Kanevski et Rimaz & Métrailler (espace) ; Goldsmith & Ermolaeva, Jacquin, Margot et Meyer (langage) ; enfin Guex et Loup, dont les contributions présentent deux manières différentes de conjuguer spatial et linguistique.

En tant que mode de structuration de l'information, les réseaux ou graphes sont également au cœur des intérêts de recherche de François Bavaud et bon nombre de ses publications explorent leurs propriétés et applications (cf. Bavaud, 2006, 2010; Bavaud & Guex, 2012; Bavaud & Métrailler, 2023; Ceré & Bavaud, 2019, ainsi que plusieurs des publications déjà mentionnées au paragraphe précédent). Dans ce volume, les contributions de Courtain & Saerens, Pante, Rimaz & Métrailler, Rochat et Rozenblat & Rogov donnent une place fondamentale aux réseaux et à leur analyse. Il est d'ailleurs remarquable que trois d'entre elles aient aussi en commun d'être centrées sur le jeu (notamment vidéo), objet d'étude en plein essor à la faculté des Lettres de l'UNIL et mettent ainsi en évidence la pertinence des réseaux pour la modélisation et l'analyse formelles des objets (vidéo-)ludiques de multiples façons.

Quiconque est familier·ère du travail de François Bavaud conviendra sans doute que la science et sa méthodologie – en particulier la démarche hypothético-déductive, envisagée comme ensemble de pratiques et procédures formelles visant à construire des modèles à partir de données, jauger des théories à l'aune de faits empiriques, sélectionner des hypothèses sur la base d'observations (Bavaud, 1998a) –

sous-tendent la majeure partie de sa recherche et son enseignement. Ce bagage méthodologique initialement acquis dans sa formation en physique et mathématiques à l'Université de Fribourg et à l'EPFL, François l'a transporté en sciences humaines et sociales (sciences sociales et politiques à l'UNIL et psychologie à l'Université de Genève), pour enfin le déposer en Lettres et Géosciences à l'UNIL, contribuant depuis lors à y développer une perspective computationnelle sur la construction des savoirs en SHS qui constitue aujourd'hui encore l'une des particularités remarquables de cette institution. Rien d'étonnant, donc, à ce que les réflexions sur la science et les corollaires institutionnels que sont les notions de disciplinarité et d'interdisciplinarité aient trouvé leur place dans les contributions qui composent cet ouvrage, en particulier celles de Jacquin, Jeanson, Pante, Piotrowski et Rozenblat & Rogov, qui offrent plusieurs points de vue sur la place des méthodes quantitatives en sciences humaines.

Espace, langage, réseaux et science sont donc, assurément, des catégories qui permettent de structurer une part importante des travaux de François Bavaud, au moins depuis sa stabilisation à l'UNIL en 2002. Du point de vue de leur adéquation à la description des quinze contributions qui forment le présent volume, en tout cas, il faut les envisager comme des catégories *floues*, dans la mesure où la plupart des contributions se rattachent à plus d'une de ces thématiques. Une telle configuration implique qu'un partitionnement *dur* des contributions ne pourrait être effectué qu'au prix d'une réduction considérable de l'information initiale, compression *avec perte* à laquelle nous n'avons pu nous résoudre. C'est donc dans l'ordre alphabétique du nom des auteur·e·s qu'elles sont finalement présentées dans ce volume et brièvement résumées dans la suite de cet avant-propos.

La contribution de Roger Bivand (Norwegian School of Economics), « Implementing a weighted measure of multivariate spatial autocorrelation », explore la mise en œuvre de la mesure pondérée d'autocorrélation spatiale multivariée δ développée par Bavaud (2024). Bivand met en avant l'importance des poids régionaux dans l'analyse de l'autocorrélation spatiale, présente une implémentation de δ en R. Il l'applique à des données électorales françaises et examine différentes

méthodes de pondération spatiale. Sa contribution valide ainsi l'approche originale en reproduisant ses résultats et en comparant δ à des mesures classiques comme l'indice de Moran.

Dans leur contribution intitulée « A bag-of-paths graph model with Poisson-distributed path lengths », Sylvain Courtain et Marco Saerens (Université catholique de Louvain) étendent le modèle « bag-of-paths » en imposant une distribution de Poisson sur la longueur des chemins dans un réseau. Cette approche permet de contrôler l'impact de la longueur des chemins, avec des applications en classification et détection de communautés. Courtain et Saerens présentent une nouvelle distance entre nœuds (*Poisson surprisal distance*), basée sur la probabilité de choisir un chemin entre deux points. Cette distance est évaluée en classification (semi-)supervisée et obtient de bonnes performances relativement aux distances existantes.

La contribution de John Goldsmith (University of Chicago) et Marina Ermolaeva (Moscow State University), « Exploring oppositions in morphology », se penche sur la question – relativement peu étudiée hors de la tradition structuraliste de la première moitié du XX^{e} siècle – des méthodes et procédures formelles permettant d'identifier les unités élémentaires qui composent les productions langagières et les systèmes linguistiques sous-jacents. En particulier, Goldsmith et Ermolaeva proposent une réflexion sur les manières de formaliser la notion de différence entre des séquences de symboles et démontrent leur utilité pour la découverte des unités morphologiques et des processus de variation allomorphique.

Dans sa contribution intitulée « A framework for spatial clustering of textual objects: applications in topic clustering and text segmentation », Guillaume Guex (UNIL) présente un formalisme statistique dont dérivent deux méthodes pour le clustering spatial d'objets textuels. Ces méthodes, aux performances compétitives relativement à l'état de l'art, équilibrent similarité sémantique et proximité des objets dans un texte pour effectuer simultanément la segmentation textuelle et l'extraction de topics. Elles constituent ainsi des outils de lecture distante pour extraire de façon non supervisée des informations thématiques à partir d'un document unique.

La contribution de Jérôme Jacquin (UNIL), « La pragmatique de corpus comme lieu d'expérimentation des méthodes mixtes : pour une interdisciplinarité focalisée », explore les développements récents de la pragmatique de corpus avec un accent particulier sur les méthodes combinant perspectives quantitative et qualitative. À travers l'exemple de l'analyse de correspondances multiples et des tests d'accords inter-annotateurs, Jacquin montre comment la statistique et les méthodes computationnelles peuvent enrichir la recherche actuelle en pragmatique et plaide en faveur d'une collaboration plus étroite entre ces disciplines.

Loïc Jeanson (UNIL) retrace, dans sa contribution intitulée « Analyse spatiale, émergence en géographie, en archéologie et en histoire », l'évolution de l'analyse spatiale et son adoption graduelle dans plusieurs disciplines. Il explore la transformation de la géographie dès les années 1950 avec l'introduction des méthodes quantitatives, marquant un tournant entre géographie descriptive et analytique. Il montre ensuite comment l'archéologie puis l'histoire se sont progressivement emparées des outils d'analyse spatiale dans le cadre du *spatial turn*, pour conclure sur l'importance croissante de ces méthodes dans les sciences humaines et sociales et le rôle central du numérique dans ces évolutions.

La contribution de Mikhail Kanevski (UNIL), « On machine learning from environmental data », propose une revue exhaustive de l'application des algorithmes de machine learning à l'analyse, la modélisation et la visualisation des données environnementales. Il présente une méthodologie complète, allant de la collecte et l'exploration des données à la sélection de variables et l'entraînement des modèles. En conclusion, il insiste sur la nécessité d'une expertise combinée en science des données et en géosciences pour exploiter efficacement le machine learning dans l'étude des phénomènes environnementaux et la prise de décision.

Dans son étude intitulée « Entre toponymes et situation géographique : cartographie et autocorrélation spatiale des suffixes communaux suisses », Romain Loup (UNIL) combine analyse spatiale et analyse linguistique pour explorer la distribution géographique des suf-

fixes dans les noms de communes suisses. Les résultats mettent en évidence une forte concentration de certaines terminaisons toponymiques dans des zones géographiques spécifiques. Cette étude propose ainsi un nouvel éclairage sur la formation des identités territoriales en Suisse et contribue à une meilleure compréhension de la diversité linguistique et culturelle nationale à travers le prisme des noms de lieux.

La contribution de Cédric Margot (UNIL), « La vilaine TINA et le capitalisme autoritaire », revisite l'argument principal de sa thèse sur le présupposé de mesure dans le discours de l'information. Margot explique que ce présupposé invalide de nombreuses affirmations reposant sur des indicateurs statistiques, en particulier celles qui, en sciences économiques, prétendent qu'il n'y a pas d'alternative (*There is no alternative – TINA*). Il critique l'utilisation des statistiques comme outils de légitimation des politiques néolibérales, propose une réflexion épistémologique sur la dépolitisation de la quantification et appelle à une prise de conscience des fondements idéologiques qui sous-tendent les propos chiffrés dans les discours médiatiques et politiques.

Dans sa contribution intitulée « πάντα ῥεῖ : changements sémantiques dans la terminologie mathématique », Robin Meyer (UNIL) retrace l'origine et le développement des termes « mathématique » et « statistique », en particulier la façon dont leurs significations ont changé au fil du temps. Il montre que le premier, qui recouvrait autrefois des disciplines comme la musique et l'astronomie, a subi une restriction sémantique, tandis que le second a vu son sens s'étendre de la gestion de l'État à l'analyse de données. La contribution met ainsi en lumière l'importance de connaître l'histoire de la terminologie pour comprendre le développement de la pensée dans des disciplines scientifiques.

La contribution d'Isaac Pante (UNIL), « Le doigt et la lune », défend la légitimité de l'informatique en Faculté des lettres, retraçant son histoire institutionnelle et soulignant son potentiel comme science humaine à part entière plutôt que simple auxiliaire méthodologique. Pante appelle la discipline à se doter d'un corpus propre et désigne le jeu (vidéo), objet culturel mêlant narration, esthétique et mécanique computationnelle, comme candidat privilégié. Il illustre cette perspec-

tive par l'étude des « livres dont vous êtes le héros », dont l'analyse via la théorie des graphes permet de croiser regard littéraire et analyse computationnelle. L'informatique y apparaît comme un espace d'indiscipline féconde, entre littérature, jeu et création.

L'essai de Michael Piotrowski (UNIL), intitulé « Les humanités numériques – vers une "mathématique originale" ? », offre une réflexion épistémologique sur le rôle des mathématiques dans les sciences humaines, arguant qu'il ne se limite pas à la quantification des phénomènes mais permet l'invention de nouvelles structures. Il retrace l'évolution de la conception des mathématiques, de l'art de la mesure à la science des modèles, et plaide en faveur d'une meilleure intégration des méthodes informatiques avec leurs fondements logiques et mathématiques. Il insiste enfin sur l'importance de définir rigoureusement les humanités numériques pour tirer pleinement parti des mathématiques comme fondement méthodologique de la construction de modèles en sciences humaines.

La contribution de Loris Rimaz et Coline Métrailler (UNIL), « "May you find peace on your journey" : une approche comparative des mondes de FromSoftware », propose une analyse formelle des mondes des jeux vidéo de FromSoftware à travers leur représentation en graphes orientés. Cette approche permet de comparer leur structure spatiale, de mesurer leur linéarité ou interconnexion et d'évaluer les expériences spatiales des joueurs. Les résultats montrent une tendance à la linéarisation progressive des mondes, tout en mettant en lumière la distinction entre espace perçu, conçu et vécu. S'appuyant sur des mesures de graphes et des visualisations, l'étude ouvre des pistes pour des analyses plus poussées avec des données dynamiques ou télémétriques.

Dans sa contribution intitulée « Des mathématiques et des jeux. Le caractère ludique des graphes et des réseaux », Yannick Rochat (UNIL) explore les liens entre jeux et mathématiques. Il présente un éventail de jeux reposant explicitement ou implicitement sur des graphes – du Dobble au Monopoly, en passant par Catan, Pandemic ou Scotland Yard – en distinguant trois approches : les jeux issus de théorèmes mathématiques, ceux intégrant des structures relationnelles et ceux qui, sans que les mathématiques soient au centre de leurs mécaniques, se

prêtent à une modélisation mathématique. Le texte met en évidence la valeur pédagogique et créative des graphes, jusqu'à leur usage dans la littérature ludique, notamment chez Perec et l'Oulipo.

Enfin, la contribution de Céline Rozenblat et Mikhail Rogov (UNIL), « Les réseaux des travaux de François Bavaud dans tous leurs étaux ou comment aborder François Bavaud dans le style de "Cantatrix Sopranica" », propose une exploration de la bibliographie de François Bavaud bien plus détaillée que la synthèse rapide et partielle faite dans cet avant-propos. Les auteur·e·s appliquent des méthodes d'analyse textuelle et d'analyse de réseaux au corpus de ses publications, ainsi qu'à ceux des travaux qu'il cite ou qui le citent. Ils mettent en évidence le rôle central de trois articles (Bavaud, 1991, 1998b, 2011) qui illustrent ses contributions sur la mécanique statistique, les matrices spatiales pondérées et les transformations de Schoenberg. Ils soulignent en particulier la créativité mathématique de François, la rigueur de sa pensée, sa capacité à construire des ponts entre des domaines variés et, par là même, à avoir un impact sur la recherche dans ces domaines.

Au terme de ce bref aperçu des belles prises ramenées dans nos filets, il nous reste à espérer qu'elles sauront susciter l'intérêt des lecteur·rice·s, et tout particulièrement celui de François Bavaud. Ce périple intellectuel à travers la modélisation de l'espace et du langage, l'analyse des réseaux et la réflexion sur la science lui est présenté comme un témoignage de notre reconnaissance pour sa recherche et son enseignement, qui constituent pour nous une source intarissable d'inspiration. Puisse cet hommage refléter l'empreinte indélébile qu'il laisse sur nos trajectoires académiques et personnelles.

Références

Bavaud, F. (1991). Equilibrium properties of the Vlasov functional: the generalized Poisson-Boltzmann-Emden equation. *Reviews of Modern Physics*, 63(1):129–149.

Bavaud, F. (1998a). *Modèles et données : une introduction à la Statistique uni-, bi- et trivariée*. L'Harmattan, Paris.

Bavaud, F. (1998b). Models for spatial weights: a systematic look. *Geographical Analysis*, 30(2):153–171.

Bavaud, F. (2005). Using local formalism in quantitative geography: a straightforward method for taking spatial auto-correlation into account. In *14th European Colloquium on Theoretical and Quantitative Geography*.

Bavaud, F. (2006). Spectral clustering and multidimensional scaling: a unified view. In Batagelj, V., Bock, H.-H., Ferligoj, A., & Ziberna, A. (éd.), *Data science and classification*, pages 131–139. Springer, Heidelberg.

Bavaud, F. (2008a). The endogenous analysis of flows, with applications to migrations, social mobility and opinion shifts. *Journal of Mathematical Sociology*, 32:239–266.

Bavaud, F. (2008b). Local concentrations. *Papers in Regional Science*, 87(3):357–370.

Bavaud, F. (2010). Euclidean distances, soft and spectral clustering on weighted graphs. In Balcázar, J. L., Bonchi, F., Gionis, A., & Sebag, M. (éd.), *Machine Learning and Knowledge Discovery in Databases*, pages 103–118, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bavaud, F. (2011). On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28:297–314.

Bavaud, F. (2013). Testing spatial autocorrelation in weighted networks: the modes permutation test. *Journal of Geographical Systems*, 15(3):233–247.

Bavaud, F. (2018). Combien y a-t-il de variétés distinctes d'anglais? *Cahiers du Centre de Linguistique et des Sciences du Langage*, 56:11–30.

Bavaud, F. (2024). Measuring and testing multivariate spatial autocorrelation in a weighted setting: A kernel approach. *Geographical Analysis*, 56(3):573–599.

Bavaud, F., Cocco, C., & Xanthos, A. (2012). Textual autocorrelation: formalism and illustrations. In *Proceedings of JADT, 11èmes Journées internationales d'analyse statistique des données textuelles*, page 109–120. Université de Liège.

Bavaud, F., Cocco, C., & Xanthos, A. (2015). Textual navigation and autocorrelation. In Mikros, G. K. & Macutek, J. (éd.), *Sequences in Language and Text*, pages 35–56. De Gruyter Mouton, Berlin, München, Boston.

Bavaud, F. & Guex, G. (2012). Interpolating between random walks and shortest paths: A path functional approach. In Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., & Guéret, C. (éd.), *Proceedings of the 4th International Conference on Social Informatics (SocInfo '12)*, volume 7710 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Bavaud, F., Kordi, M., & Kaiser, C. (2018). Flow autocorrelation: a dyadic approach. *The Annals of Regional Science*, 61(1):95–111.

Bavaud, F. & Métrailler, C. (2023). A (dis)similarity index for comparing two character networks based on the same story. In Rochat, Y., Métrailler,

C., & Piotrowski, M. (éd.), *Proceedings of the Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022)*, pages 33–42. CEUR Workshop Proceedings.

Bavaud, F., Picca, D., & Curdy, B. (2006). Non-linear correspondence analysis in text retrieval: a kernel view. In *Proceedings of JADT, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*, pages 741–747.

Bavaud, F. & Xanthos, A. (2002). Thermodynamique et statistique textuelle : concepts et illustrations. In *Proceedings of JADT, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, pages 101–111.

Bavaud, F. & Xanthos, A. (2005). Markov associativities. *Journal of Quantitative Linguistics*, 12(2-3):123–137.

Ceré, R. & Bavaud, F. (2017). Multi-labelled image segmentation in irregular, weighted networks: A spatial autocorrelation approach. In *Proceedings of the 3rd International Conference on Geographical Information Systems Theory, Applications and Management*. SCITEPRESS - Science and Technology Publications.

Ceré, R. & Bavaud, F. (2019). Soft image segmentation: On the clustering of irregular, weighted, multivariate marked networks. In Ragia, L., Laurini, R., & Rocha, J. G. (éd.), *Geographical Information Systems Theory, Applications and Management*, pages 85–109, Cham. Springer International Publishing.

Ceré, R., Egloff, M., & Bavaud, F. (2018). Geographical exploration and analysis extended to textual content. In Winter, S., Griffin, A., & Sester, M. (éd.), *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 23:1–23:7, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Cocco, C., Pittier, R., Bavaud, F., & Xanthos, A. (2011). Segmentation and clustering of textual sequences : a typological approach. In Angelova, G., Bontcheva, K., Mitkov, R., & Nicolov, N. (éd.), *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pages 427–433. RANLP 2011 Organising Committee.

Egloff, M. & Bavaud, F. (2018). Taking into account semantic similarities in correspondence analysis. In *Proceedings of the Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018)*, volume 2314 of *CEUR Workshop Proceedings*, pages 45–51.

Guex, G. & Bavaud, F. (2015). Flow-based dissimilarities: Shortest path, commute time, max-flow and free energy. In Lausen, B., Krolak-Schwerdt, S., & Böhmer, M. (éd.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, pages 101–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Guex, G., Loup, R., & Bavaud, F. (2023). Estimation of flow trajectories in a multi-lines transportation network. *Applied Network Science*, 8(44).

Loup, R. & Bavaud, F. (2025). Spatial autocorrelation of political opinions: A kernel approach. *Journal of Geographical Systems*.

Cahiers du CLSL, n° 69, 2025, pp. 19-35

Implementing a weighted measure of multivariate spatial autocorrelation

Roger Bivand
Norwegian School of Economics
Roger.Bivand@nhh.no

Abstract

Bavaud (2024) builds on and significantly broadens earlier work on measuring spatial autocorrelation, extending to multivariate settings and in particular regional weights. While measurement of spatial autocorrelation in multivariate data has been approached previously, the addition of regional weights is a major advance, as regions often differ in their contributions to global measures. Thus far, the proposed implementation described here involves dense matrices, as does much multivariate analysis. It is shown that the implementation largely reproduces the results presented in Bavaud (2024).

1 Introduction

When exploring and analyzing areal data, it is known that the definition of the areal (regional) observations may influence our results. The administrative boundaries used for data collection and aggregation are seldom chosen by the researcher. Some entities may be smaller or larger than others in area or population. Other regional structures, such as economic or social clusters, may be divided among several observations. Entitation subject to these conditions can lead to observations that are internally heterogeneous, spatially autocorrelated, and potentially heteroscedastic (Gelfand, 2010; Haining, 2010). While we cannot resolve the problem of internal heterogeneity, spatial autocorrelation can be addressed using well-known spatial statistics methods, leaving heteroscedasticity, here most likely stemming from the uneven weighting given to observations.

In Bavaud (2024), the importance of regional weights is brought into sharp focus, and measures and tests for multivariate (also univariate) spatial autocorrelation are proposed. In Bivand (2017), it was argued that modelling the Boston housing value data set without taking the number of houses used to calculate the median per spatial unit, varying from 5 to 3031, could and probably does lead to spurious inferences. So measures of spatial autocorrelation handling regional weights thoroughly are very welcome, and deserve to be implemented in software in order that they may be applied where appropriate. As I still maintain the **spdep** R package (Bivand, 2025), which provides a range of other measures of spatial autocorrelation, I chose to use R for coding the δ measure proposed in Bavaud (2024), and at this stage not to attempt to avoid the use of dense matrices.

For the remainder of this description of how the measures have been implemented, page, equation, table and figure numbers, given without a specific reference, refer to pages, equations, tables or figures in Bavaud (2024).

1.1 Data sets

The implementation of the δ measure and the methods for constructing spatial weights has been based on the data used in Bavaud (2024), that is, the toy data set provided there (p. 587), and data from the French presidential election in 2022 (pp. 589-590).

The French department boundaries (without Corsica and overseas territories) were taken from GADM, administrative level 2,[1] and simplified with a tolerance of 25m. The political data agree with the tabulation reported in Wikipedia[2], while the social data were provided by colleagues in Lausanne. The adjacency matrix also provided by colleagues in Lausanne agrees with that generated from the department boundaries.

For purposes of comparison, brief use will also be made of the Guerry data set (Anselin, 2019; Anselin & Li, 2020; Dray & Jombart, 2011; Friendly & Dray, 2023).

1 https://geodata.ucdavis.edu/gadm/gadm4.1/gpkg/gadm41_FRA.gpkg, accessed 19 Apr. 2025.

2 https://en.wikipedia.org/wiki/2022_French_presidential_election, accessed 19 Apr. 2025.

2 Regional weights

Weights may be used in statistical exploration and analysis when the observed entities are characterised by differences in their relative importance. These weights, termed *analytical*, *reliability* or *precision weights* are typically used when the observations are aggregate values constructed from differing numbers of aggregated components, such as numbers of inhabitants. Here we term these weights *regional weights*.

On page 576, the specification of the regional weights is given, requiring that the relative importance of each region be non-zero and that the regional weights sum to unity. In the 2022 French presidential election data set, the regional weights were the sums of valid votes cast for first round candidates by the regional aggregate unit, the department, divided by the sum of all valid votes cast for all mainland departments. The regional weights vary from 0.0014, (Lozère) to 0.0379, (Nord), as also seen on page 589. Uniform regional weights (p. 580), $f_i = 1/n$ where n is 94, are all 0.0106.

3 Dissimilarity matrices

Dissimilarity matrices are used in multivariate analysis to describe the relative differences between observations based on the values of the observed variables. The dissimilarity matrix used here can be constructed both for the univariate and multivariate settings; the basic specification is that $d_{ij} \geq 0$, $d_{ij} = d_{ji}$ and $d_{ii} = 0$ (Eq. 11, p. 579); the matrix should be symmetric with zeros on the leading diagonal and off-diagonal elements non-negative. Here, the term *dissimilarity matrix* is used, rather than *feature dissimilarity matrix*, as the term *feature* may mean *variable* in data science, but means *spatial entity* in much of the geospatial literature. The dissimilarity matrix is constructed from the one or more variables whose joint spatial autocorrelation is being analysed. So far, the variables used for calculating the dissimilarities between observations have been assumed to be numeric rather than categorical or a mixture of numeric and categorical.

Construction of dissimilarity matrices in multivariate analysis with numerical variables is to use Euclidean or equivalently squared Euclidean distance: $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$ (for p variables

of interest, pages 579-580 and equation 35, $\mathbf{D}$ is dense by construction), this is used for two of the French 2022 data set examples, the share of the first round vote achieved by Emmanuel Macron (data set $\mathbf{x}$), and the five social variables (data set $\mathbf{Y}$). As Anselin (2019, p. 138) points out, this is directly associated with Geary's c; Anselin (2019) also suggests that the scaling of variables of interest to mean zero and unity variance is advisable.

On page 586, extension to a squared Euclidean constructed as a chi-square dissimilarity is discussed for the vote counts for all $m = 12$ candidates in the first round of the 2022 election (data set $\mathbf{X}$). We know that the regional weights have been taken as the proportion of total votes cast in each department:

$$f_i = \frac{\sum_{k=1}^{m} x_{ik}}{\sum_{j=1}^{n} \sum_{k=1}^{m} x_{jk}}$$

The dissimilarity for votes cast by candidate is then taken as the sum over the m candidates of the difference between the vote share of the candidate between departments multiplied by the inverse of the total vote share of that candidate (equation 34, rather suggestive of shift-share analysis, e.g. Bivand, 1999):

$$d_{ij} = \sum_{k=1}^{m} \frac{\sum_{h=1}^{n} \sum_{l=1}^{m} x_{hl}}{\sum_{h=1}^{n} x_{hk}} \left(\frac{x_{ik}}{\sum_{l=1}^{m} x_{il}} - \frac{x_{jk}}{\sum_{l=1}^{m} x_{jl}} \right)^2$$

4 Constructing spatial weights

When observations may be dependent on each other in multi-level designs, time series or by being proximate neighbours in space, their mutual dependency may be expressed as a matrix or equivalently a graph. In multi-level designs, the relationships between say pupils in a school class are shown as a block-diagonal matrix, where all pupils in a class are related to each other, but pupils in different classes are unrelated - typically the principal diagonal is zero. In the time series case, the first sub-diagonal represents the first-lag relationship, between time t and $t - 1$. Cliff & Ord (1973) systematized the relationships

of proximate neighbours in space, shown as unity in the binary spatial weights matrix where observation i is a neighbour of observation j, and zero otherwise.

The introduction of regional weights means that the construction of spatial weights needs to be adjusted. Bavaud (1998) pointed out that row-standardised spatial weights constitute the conditional probability of visiting j straight from i. Such row-standardised spatial weights are frequently used, and the elements of the spatial weights matrix are set with rows summing to unity. All the regions in the adjacency matrix must be able to communicate with each other; no islands or sub-graphs are permitted. On page 577, it is stressed that the chosen regional weights (all non-zero and summing to unity) should constitute the unique stationary distribution of the weight-compatible (adjusted) spatial weights. Equation 2 on page 577 defines adjusted spatial weights as row-standardised, weight-compatible and reversible, if they satisfy:

$$\mathbf{W} \geq 0, \mathbf{W}\mathbf{1} = \mathbf{1}, \mathbf{W}^\top \mathbf{f} = \mathbf{f}, \mathbf{\Pi}\mathbf{W} = \mathbf{W}^\top \mathbf{\Pi}, \mathbf{\Pi} = \mathrm{diag}(\mathbf{f})$$

where $\mathbf{f}$ is a vector of regional weights, $\mathbf{W}$ are the adjusted spatial weights, and $\mathrm{diag}()$ is an operator creating a diagonal matrix with the given vector on the principal diagonal; $\mathbf{1}$ is an n-vector of ones.

Definition 1 on page 579 presents three ways of constructing adjusted spatial weights, followed by definition 2 on page 589 showing how to adjust a matrix of distances to meet the same criteria. At this stage, it is not obvious which method of adjustment should be chosen by the analyst. These construction methods will be described next, taking $\mathbf{A}$ as a symmetric binary adjacency matrix and $\mathbf{f}$ is a vector of regional weights.

In the proposed implementation in R, standard linear algebra is used, vectorizing all operations. For comparison, the adjacency matrix $\mathbf{A}$ for the 94 French departments has $94 \times 94 = 8836$ elements, of which 238 are non-zero.

4.1 Linearised diffusive weights

Linearised diffusive weights are discussed on page 578. The adjusted spatial weights constructed in this way also require a coefficient t:

$0 < t \leq t_1$ where $t_1 = \min(\mathbf{f}/\mathbf{r})$ and $\mathbf{r} = \mathbf{A1}$ ($\mathbf{r}$ are the row sums of $\mathbf{A}$, p. 579). Here (Eq. 7):

$$\mathbf{W}_{\text{ldw}} = \mathbf{I} - t(\mathbf{\Pi}^{-1}(\text{diag}(\mathbf{r}) - \mathbf{A}))$$

where $\mathbf{I} = \text{diag}(\mathbf{1})$ is the identity matrix.

The coefficient t may also take the value $t_2 = 1/\max(\xi)$, where ξ are the eigenvalues of the adjusted Laplacian matrix given in equation 7: $\mathbf{\Pi}^{-\frac{1}{2}}(\text{diag}(\mathbf{r}) - \mathbf{A})\mathbf{\Pi}^{-\frac{1}{2}}$ (p. 583). The count of non-zero elements in $\mathbf{W}_{\text{ldw}}$ for the French departments is 570. The Laplacian term: $\text{diag}(r) - \mathbf{A}$ corresponds to the structure matrix used in the intrinsic conditional autoregressive (ICAR) model (Held & Rue, 2010, p. 209).

4.2 Metropolis–Hastings weights

Starting from the row-standardized adjacencies $\mathbf{P} = \text{diag}(1/\mathbf{r})\mathbf{A}$, also termed the natural random walk transition matrix (p. 578), regional weights are introduced: $\mathbf{Q} = \mathbf{\Pi P}$ and the by-element minima, here taken as $\text{pmin}()$, of $\mathbf{Q}$ and its transpose are found: $\mathbf{\Gamma} = \text{pmin}(\mathbf{Q}, \mathbf{Q}^\top)$, with row sums $\gamma = \mathbf{\Gamma 1}$. Then the Metropolis–Hastings weights are (Eq. 9):

$$\mathbf{W}_{\text{mhw}} = \mathbf{\Pi}^{-1}\mathbf{E}, \mathbf{E} = \text{diag}(\mathbf{f} - \gamma) + \mathbf{\Gamma}$$

Here, as in the case of linearised diffusive weights, the principal diagonal and some other elements are non-zero, but the count of non-zero elements remains small at 562.

4.3 Iteratively fitted weights

Iteratively fitted weights first add g, a very small positive value, to the binary adjacency matrix $\mathbf{A}$, then employ iterative proportional fitting to iterate the weights matrix $\mathbf{W}_{\text{ifw}}$ to row and column margins very close to the regional weights $\mathbf{f}$. In this implementation, the `Ipfp` function from the **mipfp** package is used (Barthélemy & Suesse, 2018). As g is added to input zero elements, the output matrix $\mathbf{W}_{\text{ifw}}$ is dense, with, for the French departments, 8836 non-zero elements.

4.4 Graph distance weights

While the definition on page 589 of geographic distance weights is sufficiently clear, there is a minor discrepancy between the graph distance weights described on the same page with a maximum number of edges between the most distant nodes of 11, and that found from the graph distances found from the French department boundaries used for reproduction here. Using the `distances` function in the **igraph** package (Csárdi et al., 2024), the maximum number of edges between the most distant nodes is 12, with one most distant pair of departments being Pas-de-Calais–Pyrénées-Orientales.

Here a coefficient c is needed, where $c \in (0, c_1]$ and $c_1 = \frac{-1}{\min(\mathbf{B})}$. Taking $\mathbf{\Delta}$ as the symmetric matrix of numbers of edges separating nodes, which is by definition dense, we generate the adjusted spatial weights as (Eq. 33, p. 589):

$$\mathbf{W}_{\mathrm{gdw}} = \mathbf{1}\mathbf{1}^\top + (c\mathbf{B})\mathbf{\Pi}, \mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{\Delta}\mathbf{H}^\top, \mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{f}^\top$$

The output matrix $\mathbf{W}_{\mathrm{gdw}}$ is dense, with, for the French departments, 8835 non-zero elements.

5 Implementing Bavaud's δ and comparison with other measures

As Bavaud's δ permits the analysis of spatial autocorrelation with possibly non-uniform regional weights for multivariate data, the three main arguments – the dissimilarity matrix $\mathbf{D}$, the adjusted spatial weights matrix $\mathbf{W}$, and the vector of regional weights $\mathbf{f}$ – must be prepared first. In this implementation, the regional weights $\mathbf{f}$ may be taken from the $\mathbf{W}$ object, as the constructor functions store the regional weights used as an attribute of their output matrices, in order to attempt to ensure consistency.

Starting from equation 1 on page 576, and equation 26 on page 586, Bavaud's δ is:

$$\delta = \frac{\mathrm{tr}(\mathbf{K_D}\mathbf{K_W})}{\mathrm{tr}(\mathbf{K_D})}$$

	$\mathbf{W}_a$		$\mathbf{W}_b$		$\mathbf{W}_c$	
	δ	z	δ	z	δ	z
$\mathbf{X}$	0.96	10.51	0.75	15.59	0.55	14.21
$\mathbf{Y}$	0.94	7.78	0.67	10.88	0.42	8.94
$\mathbf{x}$	0.95	5.64	0.66	7.14	0.45	6.64

TABLE 1 – Values of Bavaud's autocorrelation index δ, together with its standardized value z.

where $\mathrm{tr}()$ is the trace of the given matrix (the sum of the principal diagonal), $\mathbf{K_D}$ is the dissimilarity kernel expressing differences in the relative position of observations in the space of the variables, and $\mathbf{K_W}$ is the spatial kernel expressing differences in the relative position of observations in geographical space. The kernels are (Eq. 1, p. 576 and Eq. 16, p. 581):

$$\mathbf{K_D} = -\frac{1}{2}\mathbf{\Pi}^{\frac{1}{2}}\mathbf{HDH}\mathbf{\Pi}^{\frac{1}{2}}, \mathbf{H} = \mathbf{I} - \mathbf{1f}^{\top}, \mathbf{\Pi} = \mathrm{diag}(\mathbf{f})$$

and (Eq. 1, p. 576 and Eq. 21, p. 583):

$$\mathbf{K_W} = \mathbf{\Pi}^{\frac{1}{2}}\mathbf{W}\mathbf{\Pi}^{-\frac{1}{2}} - \mathbf{f}^{\frac{1}{2}}(\mathbf{f}^{\frac{1}{2}})^{\top}$$

Using the toy example (pp. 586-587) and the proposed implementation of δ, we find that the output gives $\delta_A = 0.13$, $\delta_B = 0.57$, and $\delta_C = -1.00$, reproducing the original results. The standard deviates, using the moments of δ given on page 591, theorem 4, are: $z_A = 1.98$, $z_B = 1.98$, and $z_C = -2.24$, where the first two agree, but the latter does not; its probability value however does agree, $p_C = 0.025$, so z_C may be a misprint.

Next, we will try to reproduce the tabular results for the three data sets observed over the French departments; results for graph distance weights are omitted because the maximum edge counts across the graph were found to differ, so they would not be expected to agree. Table 1 reproduces the results in the original table 1 on page 582. Tables 2

	$\bar{\lambda}$	ν	κ	$\alpha(\lambda)$	$\gamma(\lambda)$
X	0.00073	3.07	0.32	7.25	56.65
Y	0.59738	2.14	0.46	7.54	58.91
x	0.00002	1.00	1.00	9.49	88.01

TABLE 2 – Spectral moments and other quantities for data sets.

	$\bar{\mu}$	Var(I)	$\alpha(\mu)$	$\gamma(\mu)$
$\mathbf{W}_a$	0.7920	0.00078	-1.54	2.79
$\mathbf{W}_b$	0.2504	0.00322	0.43	-1.10
$\mathbf{W}_c$	-0.0086	0.00486	0.28	-0.79

TABLE 3 – Spectral moments and other quantities for adjusted spatial weights.

and 3 reproduce those of both parts of the original table 2 on page 594.

Tables 4 and 5 to a certain extent reproduce the original tables 3 and 4 on page 595. Note that the values of $A(\delta)$ diverge slightly from those on page 595, as consequently do the values of p_{CF}, the Cornish-Fisher correction (Eq. 45, p. 593). In addition, the values of p_{CF} are not reported for $\mathbf{W}_b$ because the values of skewness and excess kurtosis do not fall in the domain given by Amédée-Manesme et al. (2019, p. 446, Eq. 24). The difference for p_{normal} for $\mathbf{W}_a$ and $\mathbf{Y}$ may be a difference in rounding or a misprint.

So far, the prototype implementation seems to be able to reproduce the original published output except for the skewness term, the Cornish-

	$\mathbf{W}_a$		$\mathbf{W}_b$		$\mathbf{W}_c$	
	$A(\delta)$	$\Gamma(\delta)$	$A(\delta)$	$\Gamma(\delta)$	$A(\delta)$	$\Gamma(\delta)$
X	-0.335	0.289	0.0942	-0.00952	0.0607	0.0142
Y	-0.348	0.301	0.0979	-0.00984	0.063	0.0148
x	-0.438	0.447	0.1232	-0.01396	0.0794	0.0226

TABLE 4 – Expected skewness $A(\delta)$ and excess kurtosis $\Gamma(\delta)$.

	$\mathbf{W}_a$		$\mathbf{W}_b$	$\mathbf{W}_c$	
	p_{normal}	p_{CF}	p_{normal}	p_{normal}	p_{CF}
$\mathbf{X}$	$4 \cdot 10^{-26}$	$4 \cdot 10^{-24}$	$4 \cdot 10^{-55}$	$4 \cdot 10^{-46}$	$7 \cdot 10^{-29}$
$\mathbf{Y}$	$3 \cdot 10^{-15}$	$2 \cdot 10^{-18}$	$7 \cdot 10^{-28}$	$2 \cdot 10^{-19}$	$2 \cdot 10^{-15}$
$\mathbf{x}$	$8 \cdot 10^{-09}$	$2 \cdot 10^{-11}$	$5 \cdot 10^{-13}$	$2 \cdot 10^{-11}$	$2 \cdot 10^{-09}$

TABLE 5 – One-tailed significance test of δ.

Fisher correction with very large standard deviate values, and possibly the graph distance method of constructing adjusted spatial weights.

5.1 Relationships to Moran's I

Equation 14 on page 580 asserts that δ will equal Moran's I with uniform regional weights in the univariate spatial case, tested here for row-standardized adjacencies and the vote share recorded for Emmanuel Macron by department, with squared Euclidean dissimilarities used as the dissimilarity measure. The value of Moran's I is 0.54376, equalling that of δ: 0.54376.

6 Methods for δ

Beyond the simple `print` and `summary` methods for the object returned by `spatialdelta`, the prototype implementation of δ in a forthcoming version of the R package **spdep**, several other kinds of method seemed warranted. These include `plot_spatialcoords` (Fig. 1), `plot_moran` (Figs. 2 and 3), `plot_spatialscree` (Fig. 8), `plot_factorialcoords` (Fig. 5), `plot_factorialscree` (Fig. 7), `localdelta` (Eq. 30) and `cornish_fisher` (Eq. 45).

6.1 Plots

Figure 1 reproduces the upper two panels of original figure 1, with the points proportional in size to the regional weights $\mathbf{f}$ scaled to match the range of the axes. Figure 2 shows the scree plots for $\mathbf{W}_a$, $\mathbf{W}_b$ and $\mathbf{W}_c$, as in original figure 8. Finally, figure 3 shows the Moran plots for Macron vote share for $\mathbf{W}_a$, $\mathbf{W}_b$ and $\mathbf{W}_c$; the slope coefficients shown

in the x-axis label do not match δ exactly as asserted in the caption to original figure 3.

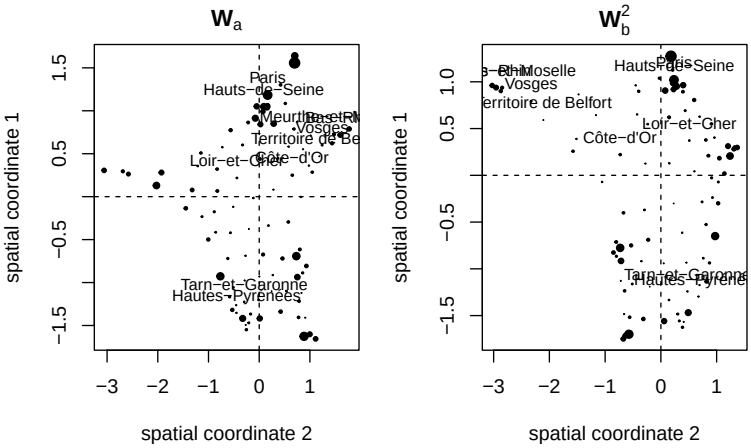


FIGURE 1 – Reproducing the upper row of original figure 1: spatial coordinates of $\mathbf{W}_a$ and $\mathbf{W}_b^2$.

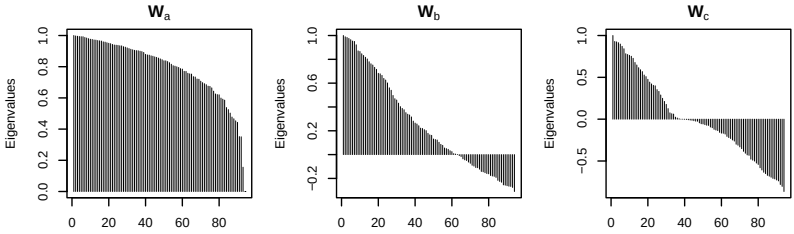


FIGURE 2 – Reproducing original figure 8: spatial scree plots.

6.2 Local δ

Local δ_i as defined in equation 30 on page 588 can be shown to equal local Moran's I_i in the Macron vote share univariate case (using squared Euclidean dissimilarities), with row-standardized adjacencies and uniform regional weights, as shown in figure 4 rendered with the **mapsf** package (Giraud, 2024). The relationship $\delta = \sum_{i=1}^{n} f_i \delta_i$ can also be

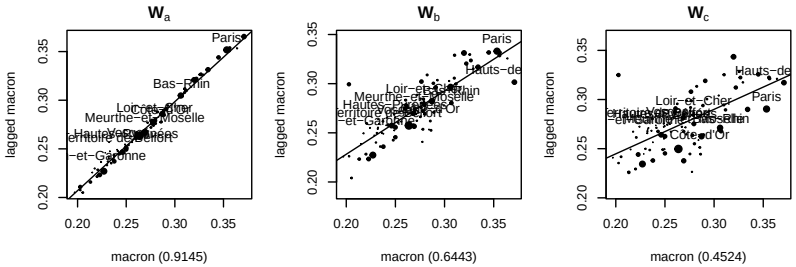


FIGURE 3 – Reproducing original figure 3: Moran-Anselin plots for Macron vote share.

shown to hold. The relationship between multivariate local δ_i and local Geary's c_i remains to be studied; figure 5 shows δ_i and local Geary's c_i for the five social variables $\mathbf{Y}$, row-standardized adjacencies and uniform regional weights.

7 Comparisons using the Guerry data set

Several explorations of spatial autocorrelation with multivariate data have used the French 85-department data set omitting Corsica, specifically six moral variables (Anselin, 2019; Anselin & Li, 2020; Dray & Jombart, 2011; Friendly & Dray, 2023).

Figures 6 and 7 show respectively the first two principal component scores of the scaled moral variables, corresponding to Anselin (2019, Figs. 1 and 2, p. 142) and Dray & Jombart (2011, Fig. 4, p. 2286) (signs of principal components flipped) which both scale the moral variables, and the first two factor coordinates of Metropolis–Hastings spatial contiguity weights of the scaled moral variables with population proportion regional weights. There is no need to show the first two factor coordinates of Metropolis–Hastings spatial contiguity weights with uniform regional weights because at least the first two are identical to the first two principal component scores of the scaled moral variables. Using population proportion regional weights, the first vector of factor coordinates is very similar to the first principal component scores, with larger differences for the second.

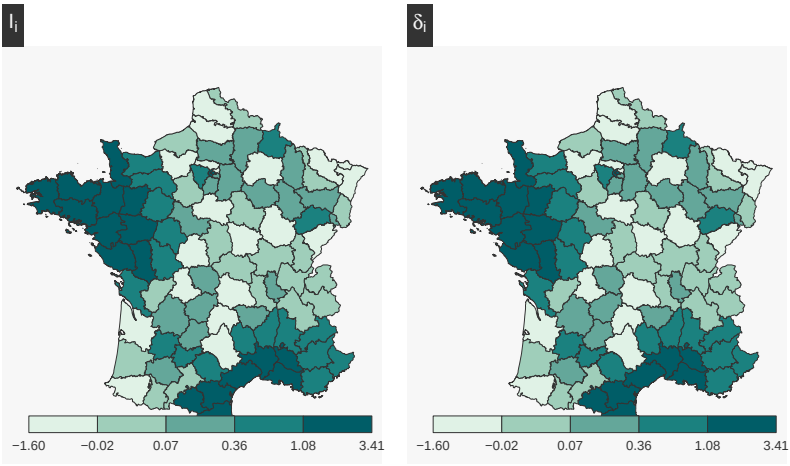


FIGURE 4 – Macron vote share with uniform regional weights and row-standardized adjacencies. Left: local Moran's I_i. Right: local δ_i.

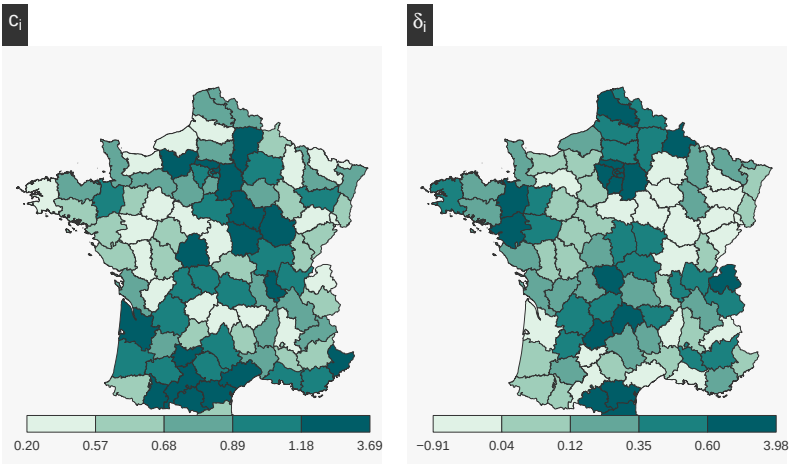


FIGURE 5 – Multivariate social data set with uniform regional weights and row-standardized adjacencies. Left: local Geary's c_i. Right: local δ_i.

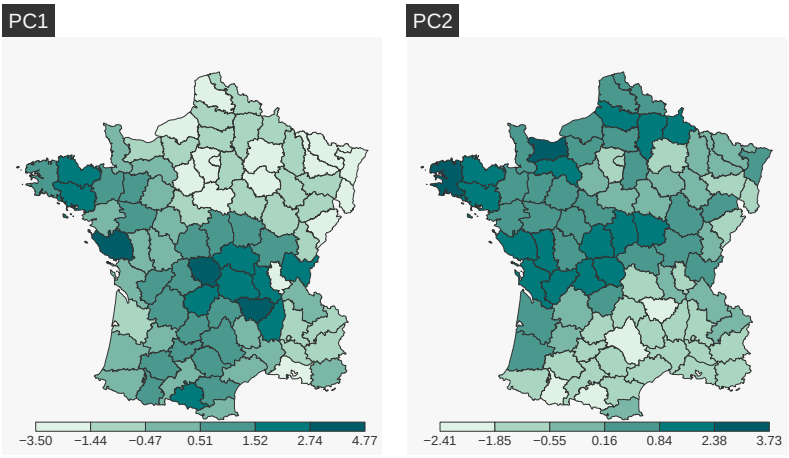


FIGURE 6 – Principal components for scaled moral variables. Left: PC1. Right: PC2.

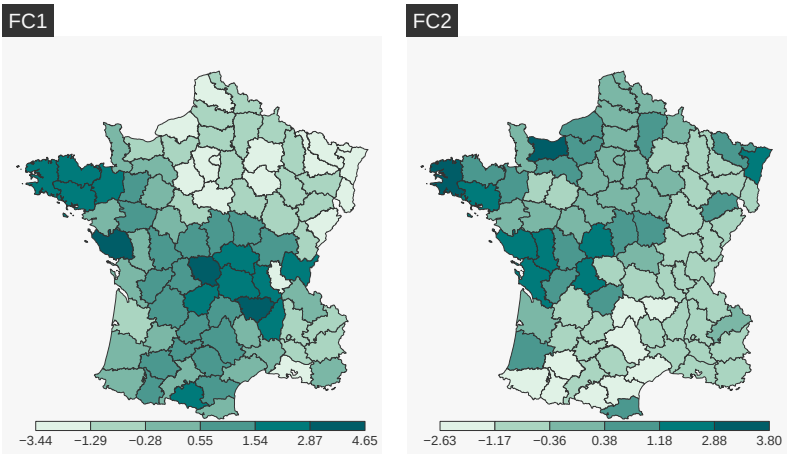


FIGURE 7 – Factorial coordinates for population weighted moral variables. Left: first factorial coordinate. Right: second factorial coordinate.

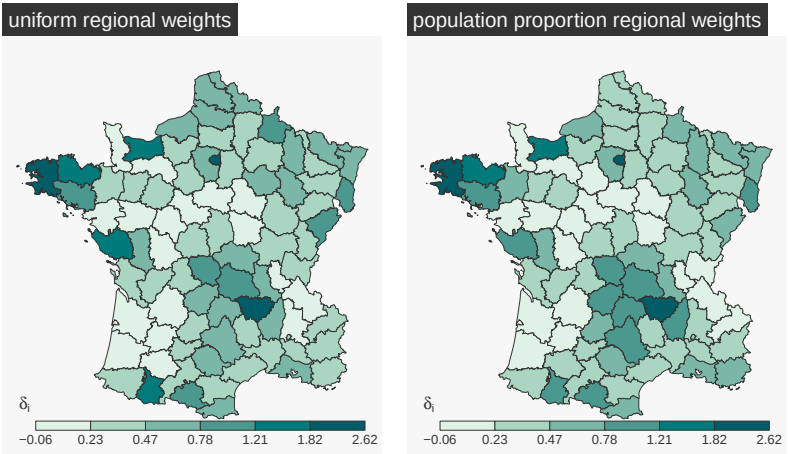


FIGURE 8 – Local δ_i values. Left: uniform regional weights. Right: population proportion regional weights.

Figure 8 shows the local δ_i values for Metropolis–Hastings spatial contiguity weights of the scaled moral variables for uniform and population proportion regional weights; here they are quite strongly correlated, but the population sizes only differ by less than an order of magnitude.

Finally, the left panel of figure 9 reproduces Anselin & Li (2020, Fig. 6, p. 502), while the right panel reproduces Anselin (2019, Fig. 11, p. 146), both approximately, as the `localC` implementation in **spdep** uses conditional permutation to calculate the pseudo-significance of local measures. Here it may well be the case that adjacency definitions vary, as Anselin (2019, footnote 6, p. 148) uses Queen contiguities, but Anselin & Li (2020, p. 496) appear to use $k = 6$ nearest neighbours, while Queen contiguities have been used here.

8 Conclusions

As reviewed in Bavaud (2024), there have been a number of suggestions going back several decades to measure spatial autocorrelation in multivariate settings. There have been some more serious studies, largely by the author himself, into the proper treatment of spatial weights, placing

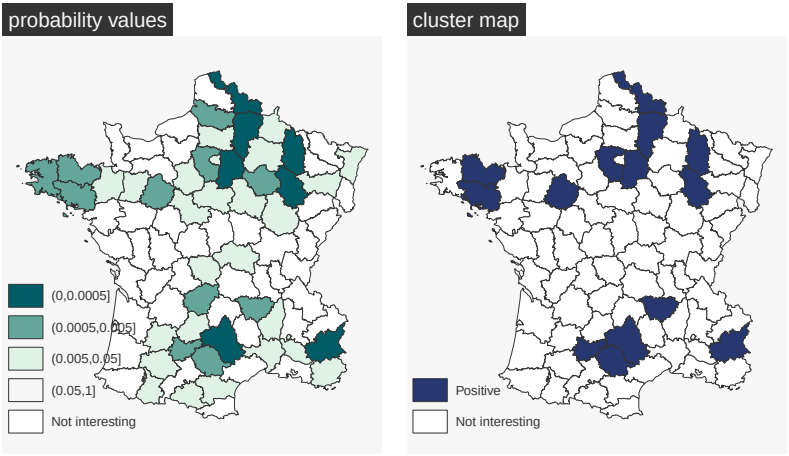


FIGURE 9 – Local c_i results. Left: classified probability values. Right: multivariate local Geary clusters.

them in the context of broader models showing how information may move between regional entities. The introduction of δ and local δ_i brings these two concerns with relationships between observations in variable space and in geographical space to a shared expression also taking into account regional weights. The implementation proposed here does need to be checked through broader use, but should assist in providing analysts with ways of exploring their data using these innovative and suggestive methods.

References

Amédée-Manesme, C.-O., Fabrice Barthélémy, & Maillard, D. (2019). Computation of the corrected Cornish–Fisher expansion using the response surface methodology: application to VaR and CVaR. *Annals of Operations Research*, 281:423–453.

Anselin, L. (2019). A local indicator of multivariate spatial association: Extending Geary's *c*. *Geographical Analysis*, 51(2):133–150.

Anselin, L. & Li, X. (2020). Tobler's law in a multivariate world. *Geographical Analysis*, 52(4):494–510.

Barthélemy, J. & Suesse, T. (2018). *mipfp: Multidimensional Iterative Proportional Fitting and Alternative Models*. R package version 3.2.1.

Bavaud, F. (1998). Models for spatial weights: a systematic look. *Geographical Analysis*, 30:153–171.

Bavaud, F. (2024). Measuring and testing multivariate spatial autocorrelation in a weighted setting: A kernel approach. *Geographical Analysis*, 56(3):573–599.

Bivand, R. (1999). Dynamic externalities and regional manufacturing development in Poland. *Tijdschrift voor Economische en Sociale Geografie*, 90(4):347–362.

Bivand, R. (2017). Revisiting the Boston data set – changing the units of observation affects estimated willingness to pay for clean air. *REGION*, 4(1):109–127.

Bivand, R. (2025). *spdep: Spatial Dependence: Weighting Schemes, Statistics*. R package.

Cliff, A. D. & Ord, J. K. (1973). *Spatial Autocorrelation*. Pion, London.

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., & Müller, K. (2024). *igraph: Network Analysis and Visualization*. R package version 2.0.3.

Dray, S. & Jombart, T. (2011). Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis. *The Annals of Applied Statistics*, 5(4):2278 – 2299.

Friendly, M. & Dray, S. (2023). *Guerry: Maps, Data and Methods Related to Guerry (1833) "Moral Statistics of France"*. R package version 1.8.3.

Gelfand, A. E. (2010). Misaligned spatial data: The change of support problem. In Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (Eds.), *Handbook of Spatial Statistics*, pages 517–539. Chapman & Hall/CRC, Boca Raton.

Giraud, T. (2024). *mapsf: Thematic Cartography*. R package version 0.12.0.

Haining, R. P. (2010). The nature of georeferenced data. In Fischer, M. & Getis, A. (Eds.), *Handbook of Applied Spatial Analysis*, pages 197–217. Springer, Heidelberg.

Held, L. & Rue, H. (2010). Conditional and intrinsic autoregressions. In Gelfand, A. E., Diggle, P. J., Fuentes, M., & Guttorp, P. (Eds.), *Handbook of Spatial Statistics*, pages 201–216. CRC Press, Boca Raton, FL.

A bag-of-paths graph framework with Poisson-distributed path lengths

Sylvain Courtain & Marco Saerens
Université catholique de Louvain
{sylvain.courtain,marco.saerens}@uclouvain.be

Abstract

This paper investigates a theoretical extension of the entropy-regularized least-cost problem on a graph from a bag-of-paths perspective. This extension constrains the a priori probability distribution on the length of the paths in order to follow a Poisson distribution. Therefore, this framework allows us to weigh the global impact of path lengths, depending on the structure of the graph, which proves useful in node classification and clustering problems. Accordingly, a novel distance measure between nodes of the graph can be defined from the probability of drawing an i-j path derived from the new bag-of-paths model. Experiments on supervised classification problems show that the proposed distance is competitive with other state-of-the-art distances and kernels on a graph.

1 Introduction

1.1 General introduction

This work aims at extending the *randomized shortest paths* (RSP) and *bag-of-paths* (BoP) models, which were introduced and refined in a series of papers (Bavaud & Guex, 2012; Kivimäki et al., 2014; Saerens et al., 2009; Yen et al., 2008), and were initially inspired by transportation science models (Akamatsu, 1996; Dial, 1971). Basically, the RSP model adds a relative entropy regularization term to the classical least-cost path problem between two nodes, i (source), j (target), of a graph, with the consequence that each i-j path is assigned a probability mass of following this path. Lower-cost paths are assigned a higher probability of being followed, although large-cost paths are less likely

to be chosen. This problem can also be viewed from the point of view of a maximum (relative) entropy problem with a fixed expected cost constraint (Saerens et al., 2009). The BoP model (Francoisse et al., 2017; Mantrach et al., 2010) generalizes the RSP model by extending the set of i-j paths to all possible paths in the graph (for all i-j pairs). As in Courtain & Saerens (2022), the present paper investigates a weighing of the i-j paths by the probability of choosing a given path length ℓ in the BoP framework. Here, the a priori path length distribution is assumed to follow a Poisson distribution, but other distributions could be used as well, depending on the application. This is interesting for at least two reasons. First, selecting paths having a certain length range allows us to quickly cover the entire network without relying on excessively long paths. For example, Backstrom et al. (2012), when analyzing a huge, popular social network, observed that the average length between two nodes was 4.74, corresponding to only 3.74 intermediaries or "degrees of separation". Second, the underlying intuition is that the Poisson distribution parameter could be seen as a resolution, scaling parameter monitoring the region of influence of each node (in terms of length from the starting node), which could prove useful in node clustering, community detection, or node label classification. However, the model introduced in Courtain & Saerens (2022) was derived in an ad hoc manner; therefore, it is reformulated in a more principled way in this paper, and applied in the experimental section to supervised classification problems.

1.2 Background and notation

Let us consider a weighted, strongly connected, directed graph G containing n nodes $\in \mathcal{V}$ (the set of nodes), and non-negative costs c_{ij} together with affinities a_{ij} (adjacency matrix) associated to (directed) edges.

1.2.1 The bag-of-paths model

The BoP model is based on the probability π_{ij} that a path $\wp$ drawn at random from a bag of paths starts at node i and ends at node j (Francoisse et al., 2017; Lebichot et al., 2014; Mantrach et al., 2010). This bag of paths is assumed to be a set containing all paths of G, of

arbitrary length. As usual, a path or walk $\wp$ is a sequence of transitions to adjacent nodes on G starting at a source node $s(\wp) = i$ and finishing at a target node $t(\wp) = j$. Moreover, the length of a path $\ell(\wp)$ is the number of hops required to follow that path. Each path is weighted according to its quality, that is, its total cost, defined as the sum of costs c_{kl} over all edges along the path $\wp$, $\tilde{c}(\wp)$. Costs associated with missing edges are supposed to be infinite, preventing these edges to be traversed.

Following Francoisse et al. (2017), two other notions need to be introduced to define the probability of drawing a path from the bag of paths. The first is the set of paths between a source node i and a target node j, including cycles, denoted by $\mathcal{P}_{ij} = \{\wp_{ij}\}$. The set $\mathcal{P}_{ij}$ usually contains an infinite, but countable, number of paths $\wp_{ij}$. The second is the set of all paths through the graph $\mathcal{P} = \bigcup_{i,j=1}^{n} \mathcal{P}_{ij}$.

In this context, the probability of drawing a path $\wp \in \mathcal{P}$ from the bag of paths, which is a probability distribution on the set $\mathcal{P}$, can be defined as the probability distribution $\mathrm{P}(\cdot)$ minimizing the total expected cost $\mathbb{E}[\tilde{c}(\wp)]$, favoring the *exploitation*, among all the distributions having a fixed relative entropy, or Kullback-Leibler divergence, J_0. The relative entropy is computed with respect to a reference distribution, here the natural random walk of the graph defining a Markov chain with transition probabilities $p_{kl}^{\mathrm{RW}} = a_{kl} / \sum_{l'=1}^{n} a_{kl'}$, allowing some random *exploration*.

The choice of this distribution naturally defines a probability distribution on the set of paths such that high-cost paths occur with a low probability while low-cost paths occur with a high probability (Francoisse et al., 2017). More precisely, we are seeking for path probabilities, $\mathrm{P}(\wp)$, $\wp \in \mathcal{P}$, minimizing the total expected cost subject to a constant relative entropy constraint,

$$\left| \begin{array}{ll} \underset{\{\mathrm{P}(\wp)\}}{\text{minimize}} & \displaystyle\sum_{\wp \in \mathcal{P}} \mathrm{P}(\wp)\tilde{c}(\wp) \\ \text{subject to} & \sum_{\wp \in \mathcal{P}} \mathrm{P}(\wp) \log\left(\mathrm{P}(\wp)/\tilde{\mathrm{P}}(\wp)\right) = J_0 \\ & \sum_{\wp \in \mathcal{P}} \mathrm{P}(\wp) = 1 \end{array} \right. \tag{1}$$

where $J_0 > 0$ is provided a priori by the user, according to the desired degree of randomness and $\tilde{\mathrm{P}}(\wp)$ represents the probability of following

the path $\wp$ (product of probabilities along the path) when walking according to the natural random walk transition probabilities p_{kl}^{RW} gathered in transition matrix $\mathbf{P}_{\text{RW}}$, and properly normalized (Francoisse et al., 2017).

1.2.2 The path-based probability distribution

Solving the problem presented in Eq. 1 leads to a *Gibbs-Boltzmann* probability distribution (see Francoisse et al., 2017, for details),

$$\mathrm{P}(\wp) = \frac{\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}}\tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{2}$$

where the parameter $\theta = 1/T$ is the inverse temperature directly related to the relative entropy J_0. Thus, as expected, low-cost paths are favored due to their high probability of being sampled. The inverse temperature parameter θ allows us to monitor the balance between exploration and exploitation. Notice that, in the sequel, it will be more convenient to provide the value of the parameter θ, with $\theta > 0$, instead of the relative entropy J_0.

Finally, the bag-of-paths probability of drawing a path starting in node $s(\wp) = i$ and ending in some other node $t(\wp) = j$ can now be defined as

$$\mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \frac{\sum_{\wp\in\mathcal{P}_{ij}}\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}}\tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{3}$$

with $\mathcal{P}_{ij}$ defining the set of all path starting at node i and ending at node j. As shown in Francoisse et al. (2017), this quantity can easily be computed in analytic closed-form, beginning with the introduction of a new transient (sub-stochastic) matrix $\mathbf{W}$ defined as

$$\mathbf{W} \triangleq \mathbf{P}_{\text{RW}} \circ \exp[-\theta\mathbf{C}] \tag{4}$$

where $\mathbf{C} = (c_{kl})$ is the cost matrix, $\mathbf{P}_{\text{RW}}$ is the natural transition probability matrix, $\circ$ is the elementwise (Hadamard) matrix product and the

exponential function is taken elementwise. Therefore, the element w_{kl} of $\mathbf{W}$ is $w_{kl} = p_{kl}^{\mathrm{RW}} \exp[-\theta c_{kl}]$.

Thanks to this matrix $\mathbf{W}$, it turns out (Francoisse et al., 2017) that the sum in the numerator of Eq. 3 can be rewritten as

$$\sum_{\wp \in \mathcal{P}_{ij}} \tilde{\mathrm{P}}(\wp) \exp[-\theta \tilde{c}(\wp)] = \sum_{\tau=0}^{\infty} \left[\mathbf{W}^{\tau}\right]_{ij} = \left[(\mathbf{I} - \mathbf{W})^{-1}\right]_{ij} = z_{ij} \quad (5)$$

with $\mathbf{I}$ being the identity matrix and where, by convention, zero-length paths are allowed and associated with a unit value and a zero cost. Thus, computing the power series of $\mathbf{W}$ leads to the definition of the matrix $\mathbf{Z} = (z_{kl}) \triangleq (\mathbf{I} - \mathbf{W})^{-1}$ called, by analogy to Markov chains (Kemeny et al., 1976), the *fundamental matrix*. Interestingly, it can be shown that elements z_{ij} can be interpreted as the expected number of times that a "killed" random walker with a transient matrix $\mathbf{W}$ starting from node i visits node j before stopping his walk (Francoisse et al., 2017). In the same way, the denominator of Eq. 3 can be computed by

$$\sum_{i,j=1}^{n} \sum_{\wp \in \mathcal{P}_{ij}} \tilde{\mathrm{P}}(\wp) \exp[-\theta \tilde{c}(\wp)] \triangleq \mathcal{Z} = \sum_{i,j=1}^{n} z_{ij} \quad (6)$$

with $\mathcal{Z}$ being the *partition function* of the bag-of-paths system.

1.2.3 The probability of drawing a path connecting two nodes

Finally, the *bag-of-paths probability* of drawing a path $\wp$ starting in node $s(\wp) = i$ and ending at some other node $t(\wp) = j$ (Francoisse et al., 2017), presented in Eq. 3, is

$$\pi_{ij} = \mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \frac{z_{ij}}{\mathcal{Z}} = \frac{z_{ij}}{\sum_{i,j=1}^{n} z_{ij}} \quad (7)$$

where $\mathbf{\Pi} = (\pi_{ij})$ is the bag-of-paths probability matrix containing the probabilities for each source-target pair of nodes. Notice that this matrix verifies $\sum_{i,j=1}^{n} \pi_{ij} = 1$ and is not symmetric in general. Therefore, in the case of an undirected graph, a variant consists of computing the probability of drawing a path $i \rightsquigarrow j$ or $j \rightsquigarrow i$, i.e. regardless of the direction of the link. The result is a symmetric matrix, $\mathbf{\Pi}_{\mathrm{sym}} = \mathbf{\Pi} + \mathbf{\Pi}^{\mathrm{T}}$,

where only the upper (or lower, since the matrix is symmetric) triangular part is relevant. The interpretation of the bag-of-paths probability matrix depends on the type of path used (regular or hitting; see Kivimäki et al., 2014 and Francoisse et al., 2017 for details); in this work, we only develop the formalism for regular paths; the case of hitting paths is left for future work.

2 Bag-of-paths with Poisson-distributed path lengths

In this section, we extend the bag-of-paths model by constraining the probability of sampling a path to follow a Poisson probability distribution on its length. Moreover, this contribution also extends the content of a previous paper (Courtain & Saerens, 2022) by avoiding the independence between path likelihood and path length. First, we introduce the BoP model constraining the probability of sampling a path to be Poisson-distributed in Subsection 2.1. Then, the joint probability of drawing a path starting in node i and ending in j is derived in Subsection 2.2. Finally, a distance measure between nodes is derived from these joint probabilities in Subsection 2.3.

2.1 BoP model with Poisson-distributed path lengths

Similarly to the standard BoP model introduced in the previous section, we minimize the free energy objective function (a reformulation of our original problem in Eq. 1), while now introducing constraints on path lengths. The idea is to constrain the probability of sampling a path $\wp$ of length $\ell(\wp) = \tau$ to follow a Poisson probability distribution $f(\tau, \lambda)$ with parameter λ.[1] This additional constraint allows us to tune the expected path length at which the relevant information can be found, as a hyper-parameter. The problem in Eq. 1 can therefore be reformulated

1 Recall the form of the Poisson distribution, $f(\tau, \lambda) = \lambda^\tau \exp(-\lambda)/\tau!$ (Papoulis & Pillai, 2002). Note that other probability distributions could also be used, depending on the problem.

as

$$\left|\begin{array}{ll} \underset{\{\mathrm{P}(\wp)\}}{\text{minimize}} & \displaystyle\sum_{i,j=1}^{n}\sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\left(\mathrm{P}(\wp)\tilde{c}(\wp)+T\,\mathrm{P}(\wp)\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}(\wp)}\right)\right) \\ \text{subject to} & \displaystyle\sum_{i,j=1}^{n}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\mathrm{P}(\wp)=f(\tau,\lambda) \quad \text{for each length } \tau \end{array}\right. \tag{8}$$

where, as before, $T=1/\theta$, $\tilde{c}(\wp)$ is the total cost of path $\wp$ when visiting the nodes in the sequential order and $\tilde{\mathrm{P}}(\wp)$ is the probability of the path $\wp$ according to the natural random walk. Furthermore, $\mathcal{P}_{ij}(\tau)$ is the set of paths connecting node i to node j whose length is exactly equal to τ.

Note that in this formulation of the problem, we do not explicitly constrain the probability distribution $\mathrm{P}(\wp)$ to sum to 1. Indeed, since the quantity $f(\tau,\lambda)$ is a Poisson probability mass, this implies that the probability distribution sums to 1, $\sum_{i,j=1}^{n}\sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\mathrm{P}(\wp)=\sum_{\tau=0}^{\infty}f(\tau,\lambda)=1$.

The problem presented in Eq. 8 can be solved by optimizing the following Lagrange function integrating equality constraints

$$\begin{aligned}\mathscr{L}(\mathrm{P}(\wp),\boldsymbol{\mu}) = &\sum_{i,j=1}^{n}\sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\left(\mathrm{P}(\wp)\tilde{c}(\wp)+T\,\mathrm{P}(\wp)\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}(\wp)}\right)\right)\\ &+\sum_{\tau=0}^{\infty}\mu_\tau\left(f(\tau,\lambda)-\sum_{i,j=1}^{n}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\mathrm{P}(\wp)\right)\end{aligned} \tag{9}$$

over the set of path probabilities $\mathrm{P}(\wp)$ with $\wp\in\mathcal{P}=\cup_{i,j=1}^{n}\cup_{\tau=0}^{\infty}\mathcal{P}_{ij}(\tau)$ (the bag of all possible paths). Minimizing Eq. 9 can be done by setting its partial derivative to the i-j path probability $\mathrm{P}(\wp)$ of length τ to zero, which gives

$$\frac{\partial\mathscr{L}(\mathrm{P}(\wp),\boldsymbol{\mu})}{\partial\mathrm{P}(\wp)}=\tilde{c}(\wp)+T\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}(\wp)}\right)+T-\mu_\tau=0 \text{ for } \wp\in\mathcal{P}_{ij}(\tau) \tag{10}$$

Isolating the logarithm and defining $\theta=1/T$ further provide

$$\mathrm{P}(\wp)=\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]\exp[\theta\mu_\tau-1] \tag{11}$$

We can now rewrite the constraint of Eq. 8, expressing that the probability of sampling a path must follow a Poisson probability distribution, from Eq. 11 as

$$f(\tau,\lambda) = \sum_{\wp\in\mathcal{P}(\tau)} \tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]\exp[\theta\mu_\tau - 1] \tag{12}$$

which means that

$$\exp[\theta\mu_\tau - 1] = \frac{f(\tau,\lambda)}{\sum_{\wp'\in\mathcal{P}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{13}$$

Finally, the i-j path probabilities $\mathrm{P}(\wp)$ of length τ can be obtained from Eqs. 11 and 13 as

$$\mathrm{P}(\wp) = \frac{f(\tau,\lambda)\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} = f(\tau,\lambda)\frac{\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\mathcal{Z}(\tau)} \tag{14}$$

where $\mathcal{Z}(\tau)$ is the partition function associated with the set of all paths with length $\ell(\wp) = \tau$, that is, $\mathcal{P}(\tau)$.

2.2 Computing the joint probability of drawing a path starting in i and ending in j

As for the standard BoP (Eq. 7), we can now define the probability of drawing a path $\wp$ starting in node $s(\wp) = i$ and ending in some other node $t(\wp) = j$, considering the set of all paths connecting i to j in exactly τ steps as $\mathcal{P}_{ij}(\tau)$. From Eq. 14, we find

$$\begin{aligned}
\pi_{ij}(\lambda) &= \mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)} \mathrm{P}(\wp) \\
&= \sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)} \frac{f(\tau,\lambda)\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{i',j'=1}^{n}\sum_{\wp'\in\mathcal{P}_{i'j'}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \\
&= \sum_{\tau=0}^{\infty} f(\tau,\lambda)\frac{\sum_{\wp\in\mathcal{P}_{ij}(\tau)} \tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{i',j'=1}^{n}\sum_{\wp'\in\mathcal{P}_{i'j'}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]}
\end{aligned} \tag{15}$$

Furthermore, we can define, in the same way as in the standard BoP model (see Eqs. 5-6 for a given τ, or Francoisse et al., 2017), the element (i,j) of the matrix $\mathbf{Z}(\tau)$ as

$$z_{ij}(\tau) \triangleq \sum_{\wp \in \mathcal{P}_{ij}(\tau)} \underbrace{\tilde{\mathrm{P}}(\wp)\exp[-\theta \tilde{c}(\wp)]}_{w(\wp)} = \sum_{\wp \in \mathcal{P}_{ij}(\tau)} w(\wp) = [\mathbf{W}^{\tau}]_{ij} \tag{16}$$

Finally, from Eq. 16, the probability of drawing a path starting in node $s(\wp) = i$ and ending in some other node $t(\wp) = j$, presented in Eq. 15, becomes

$$\pi_{ij}(\lambda) = \mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \sum_{\tau=0}^{\infty} f(\tau,\lambda)\frac{z_{ij}(\tau)}{z_{\bullet\bullet}(\tau)} \tag{17}$$

where $\bullet$ means summation on the corresponding index. We can immediately verify that $\sum_{i,j=1}^{n}\pi_{ij}(\lambda) = \sum_{\tau=0}^{\infty} f(\tau,\lambda) = 1$, as should be.

To obtain the probability $\pi_{ij}(\tau,\lambda)$ for an increasing length τ in matrix form, we now derive a recurrence expression to compute each term of the series presented in Eq. 17 in turn. To do so, from Eq. 16, we have $\mathbf{Z}(\tau) = \mathbf{W}^{\tau}$ so that we need to iterate the two following expressions until convergence,

$$\begin{cases} \mathbf{Z}(\tau+1) = \mathbf{Z}(\tau)\mathbf{W} \\ \mathbf{\Pi}(\tau+1,\lambda) = \mathbf{\Pi}(\tau,\lambda) + f(\tau+1,\lambda)\dfrac{\mathbf{Z}(\tau+1)}{z_{\bullet\bullet}(\tau+1)} \end{cases} \tag{18}$$

In this equation, $\mathbf{\Pi}(\tau,\lambda)$ contains the elements π_{ij} of the truncated series (Eq. 17) up to length τ. This update of $\mathbf{Z}(\tau)$ should usually converge quickly because spatial interactions in real-world networks are expected to be mainly local, which means that only low values of λ are relevant. In that situation, $f(\tau,\lambda)$ quickly drops to zero, which implies that the contributions to $\mathbf{\Pi}(\tau,\lambda)$ also tend to zero (each contribution in the series of Eq. 17 is $\leq f(\tau,\lambda)$). Furthermore, when $\tau = 0$, we initialize the matrices $\mathbf{Z}(\tau)$ and $\mathbf{\Pi}(\tau,\lambda)$ by

$$\begin{cases} \mathbf{Z}(0) = \mathbf{I} \\ \mathbf{\Pi}(0,\lambda) = f(0,\lambda)\dfrac{\mathbf{I}}{n} \end{cases} \tag{19}$$

with n being the number of nodes and $\mathbf{I}$ the identity matrix. This means that, for zero-length paths, the source node and the target node must be the same, and thus, $\pi_{ij}(0, \lambda) = f(0, \lambda)\delta_{ij}/n$, where δ_{ij} is the Kronecker delta.

The time complexity of the matrix $\mathbf{\Pi}(\tau, \lambda)$ computation is dominated by the matrix product performed at each iteration. Therefore, it is of order $k \cdot \mathcal{O}(n^3)$ where n is the number of nodes and k is the number of required iterations. However, complexity could be lower when working with sparse matrices.

2.3 A derived distance measure between nodes

We now derive a distance measure between the nodes following the same procedure as in Courtain & Saerens (2022) and Francoisse et al. (2017). More specifically, we take minus the (elementwise) logarithm of the probability matrix $\mathbf{\Pi}(\tau, \lambda)$ obtained after convergence of Eq. 18. The resulting distance matrix between nodes, called the *directed Poisson surprisal distance*, is defined as $\mathbf{\Delta}^{\text{DPSURP}} = -\log \mathbf{\Pi}(\tau, \lambda)$ with the log being the natural elementwise logarithm. It computes the "surprisal" of observing a path starting in i and ending in j, and is an extension of the surprisal distance introduced in Francoisse et al. (2017). The *Poisson surprisal distance* (PSurp) measure is the symmetrized quantity, $\mathbf{\Delta}^{\text{PSURP}} = \frac{1}{2}\left(\mathbf{\Delta}^{\text{DPSURP}} + (\mathbf{\Delta}^{\text{DPSURP}})^{\text{T}}\right)$, where the diagonal elements are then set to zero by subtracting $\mathbf{Diag}(\mathbf{\Delta}^{\text{PSURP}})$ to $\mathbf{\Delta}^{\text{PSURP}}$. It measures both proximity (low cost) and reachability (high connectivity) of the nodes of G; in other words, it quantifies the ease of accessibility between pairs of nodes.

3 Experiments

In this section, we compare the performance of the Poisson surprisal distance introduced in the previous section with other state-of-the-art methods regarding classification accuracy on a graph-based kernel semi-supervised classification task. It is important to emphasize that our description follows the framework established in Courtain (2022) and Courtain & Saerens (2022), and that the compared methods, experimental setup, and datasets analyzed are similar to those used in that study.

Therefore, we will only focus on the new points in our description and refer to those papers for more details.

3.1 Investigated state-of-the-art methods

As part of our experimental design, we selected three dissimilarity measures and seven kernel matrices as baseline methods to evaluate the performance of the introduced distance. These methods achieved the highest performance in their respective categories in the semi-supervised classification experiment described in Courtain (2022). Furthermore, the majority of these methods have already demonstrated strong performance in prior semi-supervised tasks (Courtain et al., 2023; Courtain & Saerens, 2022; Francoisse et al., 2017; Ivashkin & Chebotarev, 2022; Leleux et al., 2021), as well as in unsupervised tasks (Courtain et al., 2021; Ivashkin & Chebotarev, 2017; Sommer et al., 2017; Yen et al., 2009).

A summary of all the investigated methods and their acronyms is provided in table 1; notice that our proposed method is called PSurp. We refer the interested reader to subsection 7.1 and table 7.1 in Courtain, 2022 for an in-depth description of each method and the parameter values used. For our proposed method, we employed the same parameters as PWSurp (introduced in Courtain & Saerens, 2022), specifically $\theta = \{10^{-6}, 10^{-5}, \ldots, 10\}$ and $\lambda = \{1, 2, 3, 5, 10\}$. Furthermore, for PWSurp, we employed both uniform priors (PWSurpUni) and L_1-normalized degree priors (PWSurpDegree) to enable direct comparison with PSurp, even though previous studies have shown that L_1-normalized priors yield superior performance (Courtain, 2022; Courtain & Saerens, 2022).

3.2 Experimental design

To evaluate node classification performance across the 14 network datasets, we transform all dissimilarity and similarity measures into kernel matrices, $\mathbf{K}$, by removing negative eigenvalues and feeding these matrices into a kernel SVM[2] with various margin parameter values $c = \{10^{-2}, 10^{-1}, 1, 10, 100\}$. As noted earlier, the same dataset

2 The LIBSVM library (Chang & Lin, 2011) was used with options '-s 0' and '-t 4'.

Method	Acronym
Poisson surprisal distance (this paper)	PSurp
Poisson-weighted surprisal distance with priors (by Uniform and Degree distribution) (Courtain & Saerens, 2022)	PWSurp
Margin-constrained bag-of-hitting-paths surprisal distance (Guex et al., 2019)	cBoPH
Correlation kernel based on the number of occurrences of nodes on regular paths (Guex et al., 2021)	nCor
Correlation kernel based on the number of occurrences of nodes on hitting paths (Guex et al., 2021)	nCorH
Logarithmic forest distance (Chebotarev, 2011)	LF
Modified regularized Laplacian kernel (Ito et al., 2005)	MRL
Sigmoid commute time similarity (Yen et al., 2007, 2009)	SCT
Sigmoid corrected commute time similarity (based on von Luxburg et al., 2010, Yen et al., 2007, 2009)	SCCT
Logarithmic communicability similarity (Ivashkin & Chebotarev, 2017)	LogCom
Random walk with restart similarity (Tong et al., 2006)	RWWR

TABLE 1 – The different methods for computing similarities and dissimilarities between nodes investigated in our experiments, with their acronym.

collection as in Courtain (2022) is employed; detailed descriptions can be found in table 7.3 of that work. This collection comprises nine subsets from the 20 Newsgroup datasets (Lichman, 2013; Yen et al., 2009), four WebKB datasets (Macskassy & Provost, 2007), and the IMDB dataset (Macskassy & Provost, 2007). All networks are undirected, represented by an adjacency matrix $\mathbf{A}$, with transition costs c_{ij} defined as the inverse of affinity, $1/a_{ij}$, akin to electrical networks (Francoisse et al., 2017). Each node is labeled with a class for classification purposes based on the structure of the network.

For dissimilarity measures, we examine three transformations to a kernel matrix: classical multidimensional scaling (MDS)[3] (Borg & Groenen, 1997), Gaussian transformation (Gauss) (Schölkopf & Smola, 2002), and centered Gaussian transformation (GaussCenter).[4] To maintain clarity, we present only the best kernel transformation results for each method, as determined by Nemenyi tests (Demšar, 2006) (see Tab. 2 for details).

3 Kernels generated through this transformation are inherently centered.

4 In this transformation, the kernel $\mathbf{K}$ is centered as $\mathbf{K} \leftarrow \mathbf{HKH}$, where $\mathbf{H} = \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}/n$ is the centering matrix, $\mathbf{e}$ is a vector of ones, $\mathbf{I}$ is the identity matrix, and n is the number of nodes.

To minimize variability in the results, we conduct 10 repetitions of a 5×5 nested cross-validation procedure, with different labeled and unlabeled node splits in each run. In the external 5-fold cross-validation, 20% of the node labels are used for training, while the remaining 80% are hidden for testing. During each internal 5-fold cross-validation, parameters are tuned on the external training fold using 80% of the labeled nodes. External and internal folds are kept consistent across all methods in a given run to maintain comparability. The final results, presented in table 2, are the average of 50 accuracy scores from the external cross-validation folds.

3.3 Results and discussion

The results of the semi-supervised classification experiments are presented in table 2. To enhance clarity, the highest accuracy for each dataset is indicated in bold. As shown in table 2, LF outperforms all others on the WebKB datasets, while the nCor method achieves the best performance on the IMDB dataset. The results for the Newsgroups datasets are more varied: PWSurpDegree and SCCT each deliver the highest accuracy on three of these datasets, whereas PSurp, PWSurpUni, and cBoPH each attain the highest accuracy on one dataset. Overall, these results suggest that the best-performing method varies depending on the dataset.

To further investigate the results, we first conducted a nonparametric Friedman-Nemenyi statistical test, followed by multiple Wilcoxon signed-rank tests with a 95% confidence level ($\alpha = 0.05$) (Demšar, 2006) based on the average accuracy computed on the 14 datasets. The Friedman test yielded a p-value of 2.5×10^{-12}, which is below the α threshold, indicating that at least one method's performance is significantly different from the others. Given the positive result of the Friedman test, we proceeded with the Nemenyi test, with the results shown in Figure 1. To refine our analysis and provide deeper insights into the relative performance of the methods, we also conducted multiple Wilcoxon signed-rank tests, the results of which are summarized in table 3.

Dataset	PSurp	PWSurpUni	PWSurpDeg	CBoPH	nCor	nCorH
WebKB-Texas	77.37±3.94	78.44±3.14	78.55±3.13	76.68±3.26	67.36±2.47	68.01±2.10
WebKB-Washington	70.41±1.99	70.42±2.02	71.85±2.04	70.12±2.19	65.84±0.97	65.71±1.23
WebKB-Wisconsin	77.31±2.19	77.08±2.46	77.95±2.14	76.42±2.42	73.74±1.85	73.70±1.78
WebKB-Cornell	59.47±2.91	59.56±2.61	59.61±2.59	58.79±3.38	52.89±3.63	53.58±3.51
IMDB	77.57±1.36	77.52±1.47	77.95±1.33	79.24±1.48	**79.63±1.16**	79.60±1.10
Newsgroup-2cl-1	96.21±1.13	96.36±1.40	96.13±1.12	95.63±0.88	95.51±1.09	95.32±1.18
Newsgroup-2cl-2	**92.76±1.64**	92.50±1.82	92.41±1.78	92.59±1.60	92.47±1.57	92.56±1.72
Newsgroup-2cl-3	96.53±0.95	96.48±1.02	96.44±0.99	**96.62±0.74**	95.93±1.28	96.05±1.23
Newsgroup-3cl-1	93.37±1.22	93.39±1.12	93.56±1.06	93.33±1.00	92.98±1.43	92.84±1.26
Newsgroup-3cl-2	93.34±1.18	93.32±1.11	93.23±1.05	93.23±1.00	92.39±1.06	92.28±1.14
Newsgroup-3cl-3	92.74±1.23	92.78±1.35	**93.28±1.32**	93.26±1.02	92.43±1.35	92.45±1.27
Newsgroup-5cl-1	89.15±0.99	**89.37±1.02**	89.25±1.07	88.94±0.89	87.87±1.49	87.93±1.54
Newsgroup-5cl-2	84.12±1.50	84.38±1.42	**84.46±1.52**	83.94±1.32	82.78±1.42	82.76±1.42
Newsgroup-5cl-3	83.80±1.83	83.67±1.91	**84.15±1.51**	83.54±1.32	82.36±1.28	82.49±1.30

Dataset	LF	MRL	SCT	SCCT	LogCom	RWWR
WebKB-Texas	**79.46±2.72**	49.30±1.95	76.54±3.49	76.80±3.01	79.04±2.47	51.08±5.68
WebKB-Washington	**71.87±1.99**	64.95±1.39	69.49±2.49	68.66±2.18	71.35±2.41	65.36±1.36
WebKB-Wisconsin	**79.48±1.95**	50.12±0.91	77.70±2.29	77.64±2.03	78.44±2.26	63.11±4.35
WebKB-Cornell	**59.96±2.65**	41.91±0.09	58.92±2.89	59.13±2.94	58.85±2.88	40.22±5.65
IMDB	79.19±1.36	77.82±2.42	78.03±1.65	77.41±1.43	77.29±1.64	79.05±1.10
Newsgroup-2cl-1	94.95±1.72	94.46±1.64	95.84±1.03	**97.07±0.67**	95.80±1.31	94.42±1.76
Newsgroup-2cl-2	92.27±1.54	91.70±1.70	92.22±1.41	92.50±1.29	92.32±1.55	91.98±1.81
Newsgroup-2cl-3	95.53±1.56	95.02±1.26	96.00±1.00	96.06±0.75	95.27±1.59	95.65±1.19
Newsgroup-3cl-1	92.70±1.13	91.91±1.51	93.43±1.02	**93.94±0.63**	93.28±1.11	92.74±1.19
Newsgroup-3cl-2	92.35±1.33	91.52±1.50	92.25±1.15	**93.36±0.85**	92.36±1.00	92.29±1.36
Newsgroup-3cl-3	92.36±1.43	91.26±1.56	92.64±1.12	93.11±0.78	91.61±1.19	91.26±1.73
Newsgroup-5cl-1	87.95±1.17	86.26±1.67	86.96±1.15	88.00±1.03	87.77±1.46	87.48±1.26
Newsgroup-5cl-2	82.69±1.84	80.94±2.03	81.86±1.59	82.96±0.85	83.16±1.51	82.77±1.43
Newsgroup-5cl-3	82.50±1.72	80.94±1.71	82.17±1.77	83.56±1.18	82.73±1.79	82.18±1.52

TABLE 2 – Classification accuracy in percentage terms (mean ± standard deviation) for the various classification methods across different datasets. For each dataset and method, the final accuracy is computed following the experimental design outlined in Subsection 3.2. The best-performing method for each dataset is highlighted in bold.

The Nemenyi test indicates that MRL and RWWR perform significantly worse than PSurp, PWSurpUni, PWSurpDegree, cBoPH, and SCCT. Additionally, MRL is also outperformed by LF and LogCom. The test further shows that PWSurpDegree performs significantly better than SCT, nCor, and nCorH.

In addition to confirming the findings of the Nemenyi test, the Wilcoxon tests (see Tab. 3) reveal that PWSurpDegree outperforms all the evaluated methods except for LF and PWSurpUni. The tests further indicate that PSurp, PWSurpUni, cBoPH, and SCCT outperform nCor, nCorH, and SCT. Moreover, they also show that MRL performs worse than all other methods in the analysis. Finally, RWWR is outperformed by all methods, except for SCT and MRL.

Overall, the experiments demonstrated that the newly introduced dis-

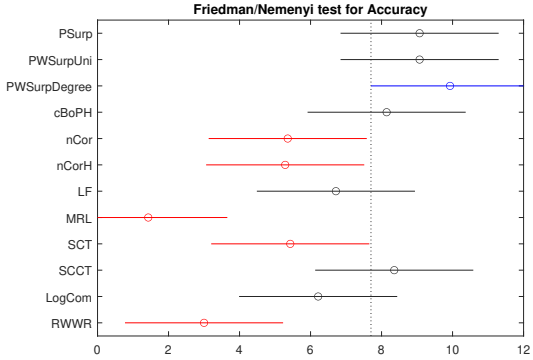


FIGURE 1 – Mean ranks and 95% Nemenyi confidence intervals for the 12 methods evaluated across 14 datasets. Significant differences between methods are determined by non-overlapping confidence intervals. The x-axis represents the average rank of each method, where a higher rank indicates better performance. The top-performing method (PWSurpDegree) and the lowest-ranked methods (nCor, nCorH, MRL, SCT, and RWWR) are highlighted.

Kernel → **Kernel ↓**	PSurp	PWSurpUni	PWSurpDegree	cBoPH	nCor	nCorH	LF	MRL	SCT	SCCT	LogCom	RWWR
PSurp	1.0000	0.7148	**0.0166**	0.0906	**0.0052**	**0.0052**	1.0000	**0.0002**	**0.0052**	0.2676	0.1726	**0.0023**
PWSurpUni	0.7148	1.0000	0.1041	0.1531	**0.0052**	**0.0067**	0.7148	**0.0002**	**0.0134**	0.2958	0.0906	**0.0023**
PWSurpDegree	**0.0166**	0.1041	1.0000	**0.0295**	**0.0040**	**0.0040**	0.2166	**0.0001**	**0.0002**	**0.0353**	**0.0040**	**0.0012**
cBoPH	0.0906	0.1531	**0.0295**	1.0000	**0.0009**	**0.0006**	0.8552	**0.0001**	**0.0295**	0.6698	0.5016	**0.0001**
nCor	**0.0052**	**0.0052**	**0.0040**	**0.0009**	1.0000	0.8077	0.7609	**0.0001**	0.3575	**0.0052**	0.2412	**0.0001**
nCorH	**0.0052**	**0.0067**	**0.0040**	**0.0006**	0.8077	1.0000	0.8077	**0.0001**	0.4263	**0.0085**	0.2166	**0.0006**
LF	1.0000	0.7148	0.2166	0.8552	0.7609	0.8077	1.0000	**0.0001**	0.0580	0.9515	0.2676	**0.0031**
MRL	**0.0002**	**0.0002**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	1.0000	**0.0001**	**0.0002**	**0.0006**	**0.0107**
SCT	**0.0052**	**0.0134**	**0.0002**	**0.0295**	0.3575	0.4263	0.0580	**0.0001**	1.0000	**0.0353**	0.2958	0.0676
SCCT	0.2676	0.2958	**0.0353**	0.6698	**0.0052**	**0.0085**	0.9515	**0.0002**	**0.0353**	1.0000	0.3910	**0.0031**
LogCom	0.1726	0.0906	**0.0040**	0.5016	0.2412	0.2166	0.2676	**0.0006**	0.2958	0.3910	1.0000	**0.0166**
RWWR	**0.0023**	**0.0023**	**0.0012**	**0.0001**	**0.0001**	**0.0006**	**0.0031**	**0.0107**	0.0676	**0.0031**	**0.0166**	1.0000

TABLE 3 – The p-values are reported from pairwise Wilcoxon signed-rank tests applied to the results presented in table 2. The p-values below the threshold of 0.05 are highlighted to indicate statistical significance.

tance achieved strong performance in our experimental setup. It proved to be competitive with cBoPH and SCCT, which have consistently shown good results across a wide range of experiments in previous semi-supervised (Courtain et al., 2023; Courtain & Saerens, 2022; Francoisse et al., 2017; Ivashkin & Chebotarev, 2022; Leleux et al., 2021) and unsupervised tasks (Courtain et al., 2021; Ivashkin & Chebotarev, 2017; Sommer et al., 2017; Yen et al., 2009).

Additionally, in line with the findings of Courtain & Saerens (2022), PWSurpDegree once again emerges as the most effective method. How-

ever, unlike PWSurpUni and PSurp, this approach incorporates priors, which could explain the observed differences in performance. Notably, experiments have demonstrated that both methods (PSurp and PWSurp) show very similar performance levels when using uniform priors. This observation was previously emphasized in an exploratory analysis, which revealed that these two distances have a correlation exceeding 90% (see Subsection 6.3.4 in Courtain, 2022).

4 Conclusions and future work

This paper investigated a mechanism constraining the probability of drawing a path from a bag of paths to follow a predefined discrete probability distribution on their length (illustrated here with a Poisson distribution). Consequently, the marginal probability of selecting a path of a given length follows this specified distribution. The introduction of this path-length distribution extends the basic BoP framework by allowing for a more precise tuning of the model, according to the application under study.

More precisely, an algorithm computing the probability of drawing a path connecting a given source and a given target node is developed. Then, taking minus the logarithm of this probability provides a dissimilarity measure between the two nodes, in terms of accessibility in the network, called the Poisson surprisal distance. This dissimilarity measure is then investigated in an experimental comparison with other state-of-the-art algorithms. The introduced measure was shown to provide competitive results.

Future work will consider other path length probability distributions (instead of Poisson), but also the introduction of a priori probabilities at source and target nodes, which showed superior results in previous work (Courtain & Saerens, 2022). In addition, it would also be interesting to compute the expected cost between two nodes within the same formalism, which would also provide a new dissimilarity measure between nodes.

Acknowledgments

We thank Professor François Bavaud of the University of Lausanne (Switzerland) and an anonymous reviewer for their comments and suggestions.

References

Akamatsu, T. (1996). Cyclic flows, Markov process and stochastic traffic assignment. *Transportation Research B*, 30(5):369–386.

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, page 33–42, New York, NY, USA. Association for Computing Machinery.

Bavaud, F. & Guex, G. (2012). Interpolating between random walks and shortest paths: A path functional approach. In Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., & Guéret, C. (Eds.), *Proceedings of the 4th International Conference on Social Informatics (SocInfo '12)*, volume 7710 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Borg, I. & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. Springer, New York, NY, USA.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Chebotarev, P. (2011). A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discrete Applied Mathematics*, 159(5):295–302.

Courtain, S. (2022). *Essays on network data analysis through the bag-of-paths framework*. PhD thesis, Université catholique de Louvain, Belgium.

Courtain, S., Guex, G., Kivimäki, I., & Saerens, M. (2023). Relative entropy-regularized optimal transport on a graph: a new algorithm and an experimental comparison. *International Journal of Machine Learning and Cybernetics*, 14(4):1365–1390.

Courtain, S., Leleux, P., Kivimäki, I., Guex, G., & Saerens, M. (2021). Randomized shortest paths with net flows and capacity constraints. *Information Sciences*, 556:341–360.

Courtain, S. & Saerens, M. (2022). A simple extension of the bag-of-paths model weighting path lengths by a Poisson distribution. In *Complex Networks & Their Applications X*, pages 220–233, Cham. Springer International Publishing.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Dial, R. (1971). A probabilistic multipath assignment model that obviates path enumeration. *Transportation Research*, 5:83–111.

Francoisse, K., Kivimäki, I., Mantrach, A., Rossi, F., & Saerens, M. (2017). A bag-of-paths framework for network data analysis. *Neural Networks*, 90:90–111.

Guex, G., Courtain, S., & Saerens, M. (2021). Covariance and correlation kernels on a graph in the generalized bag-of-paths formalism. *Journal of Complex Networks*, 8(6).

Guex, G., Kivimäki, I., & Saerens, M. (2019). Randomized optimal transport on a graph: framework and new distance measures. *Network Science*, 7(1):88–122.

Ito, T., Shimbo, M., Kudo, T., & Matsumoto, Y. (2005). Application of kernels to link analysis. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '05)*, pages 586–592.

Ivashkin, V. & Chebotarev, P. (2017). Do logarithmic proximity measures outperform plain ones in graph clustering? In Kalyagin, V. A., Nikolaev, A. I., Pardalos, P. M., & Prokopyev, O. A. (Eds.), *Models, algorithms, and technologies for network analysis*, pages 87–105, Cham. Springer International Publishing.

Ivashkin, V. & Chebotarev, P. (2022). Dissecting graph measure performance for node clustering in LFR parameter space. In *Proceedings of the 10th International Conference on Complex Networks and their Applications (CNA '21)*, pages 328–341. Springer.

Kemeny, J. G., Snell, J. L., & Knapp, A. (1976). *Denumerable Markov chains*. Springer, New York, NY, USA.

Kivimäki, I., Shimbo, M., & Saerens, M. (2014). Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A: Statistical Mechanics and its Applications*, 393:600–616.

Lebichot, B., Kivimäki, I., Francoisse, K., & Saerens, M. (2014). Semi-supervised classification through the bag-of-paths group betweenness. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1173–1186.

Leleux, P., Courtain, S., Guex, G., & Saerens, M. (2021). Sparse randomized shortest paths routing with Tsallis divergence regularization. *Data Mining and Knowledge Discovery*, 35:986–1031.

Lichman, M. (2013). UCI machine learning repository.

Macskassy, S. A. & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983.

Mantrach, A., Yen, L., Callut, J., Francoise, K., Shimbo, M., & Saerens, M. (2010). The sum-over-paths covariance kernel: A novel covariance between nodes of a directed graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1112–1126.

Papoulis, A. & Pillai, S. U. (2002). *Probability, random variables and stochastic processes*. McGraw-Hill, New York, NY, USA, 4th edition.

Saerens, M., Achbany, Y., Fouss, F., & Yen, L. (2009). Randomized shortest-path problems: Two related models. *Neural Computation*, 21(8):2363–2404.

Schölkopf, B. & Smola, A. (2002). *Learning with kernels*. MIT Press, Cambridge, MA, USA.

Sommer, F., Fouss, F., & Saerens, M. (2017). Modularity-driven kernel k-means for community detection. In Lintas, A., Rovetta, S., Verschure, P. F., & Villa, A. E. (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 423–433, Cham. Springer International Publishing.

Tong, H., Faloutsos, C., & Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM '06)*, pages 613–622.

von Luxburg, U., Radl, A., & Hein, M. (2010). Getting lost in space: Large sample analysis of the commute distance. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '10)*, pages 2622–2630. MIT Press.

Yen, L., Fouss, F., Decaestecker, C., Francq, P., & Saerens, M. (2007). Graph nodes clustering based on the commute-time kernel. In Zhou, Z.-H., Li, H., & Yang, Q. (Eds.), *Advances in Knowledge Discovery and Data Mining*, pages 1037–1045, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yen, L., Fouss, F., Decaestecker, C., Francq, P., & Saerens, M. (2009). Graph nodes clustering with the sigmoid commute-time kernel: A comparative study. *Data & Knowledge Engineering*, 68(3):338–361.

Yen, L., Saerens, M., Mantrach, A., & Shimbo, M. (2008). A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 785–793, New York, NY, USA. Association for Computing Machinery.

Cahiers du CLSL, n° 69, 2025, pp. 57-71

Exploring oppositions in morphology

John A. Goldsmith & Marina Ermolaeva
University of Chicago; Moscow State University
goldsmith@uchicago.edu; marinkaermolaeva@gmail.com

Abstract

In this paper, we define a group structure over strings and note that by applying this computation to words, we obtain major steps towards a method for identifying allomorphy and learning morphophonemics. First order differences among a set of words forming a paradigm identify *morphs*, while second order differences identify *allomorphy*. When this allomorphy appears at morpheme boundary, this can in a wide range of cases be identified as *morphophonology*.

1 Introduction

For those who delve deeply into it, language, like music, seems to reveal complex patterns of varying scales. What is the language of patterns? At its most abstract, the language is transformations, and at its more concrete, it is differences and samenesses. The student who learns a new European language has to learn various verbal paradigms, with the understanding that once we learn to conjugate *parler*, for example, that knowledge will directly extend to a large number of other verbs, because the differences among the various inflected forms of *parler* are exactly matched by the differences among the inflected forms of *sauter*, and many, many other verbs. Our goal in the work we describe here is based on the belief that by carefully building a series of notions of difference that are useful in describing language, we may achieve greater insight into the structure of language. If we let ourselves be inspired by the calculus, we are led to ask whether our notions of difference can be extended to second-order differences, just as first

derivatives are extended to second derivatives. This paper describes certain initial steps along the way to that end, and is part of a larger project in progress.[1]

2 Strings, and a group structure for multisets

Discussions of formal languages often begin by assuming an alphabet and a semigroup formed from the alphabet with concatenation, and just as often an identity element is assumed, forming a monoid. To talk about particular languages, it is necessary in addition to include the notion of *set* (or, as we will see, something *like* the notion of set, such as a *multiset*). A regular language can be defined as a semiring with two operations, *set union* and *string concatenation*, defined both with respect to pairs of strings and to sets of strings (with closure under both operations as well as closure under Kleene star operation, and complementation, though we do not discuss these latter two properties).

In this paper we are interested, however, in exploring the notion of *difference* of strings, and of sets of strings, and this requires that an *inverse* be available for at least one of the operations. In the context of arithmetic, the difference between two numbers is typically defined in a way that takes advantage of the existence of the inverse of every number under addition. The difference of 5 and 3 is 5 + (-3), and the difference of 3 and 5 is 3 + (-5). But more generally, the notion of the difference of two elements a and b can be formalized in the context of a group as the operation of the group on a and b^{-1} (the inverse of b with respect to the operation of the group).

When we turn to the semiring of strings, we find (to our surprise) that we have an embarrassment of riches: we can define inverse elements *either* with respect to the concatenation operator, *or* with respect to the set union operator – if we replace the sets of our ring by multisets. The inverse elements with respect to concatenation we must introduce in any event, but its use can be restricted to a smaller set of cases if we

1 We are pleased to dedicate this paper to François Bavaud, in the belief that our efforts to learn more about the nature of natural language through mathematical formalization will resonate with his own interest in exploring many aspects of natural language through mathematics over the course of a long and productive career.

introduce the use of multisets. This leads us to more than one solution for some of the areas which we explore, and we are led to compare the approaches to the problems we consider.[2]

We will briefly outline these two accounts. In the account in which the elements of the alphabet have inverses with respect to the operation of *concatenation*, we do not need multisets; we can restrict our attention to traditional sets. Each element of the alphabet a_i is allowed an inverse element which we denote a_i^{-1}. Thus $walkingg^{-1}n^{-1}i^{-1} = walk$, and $(the)^{-1} = e^{-1}h^{-1}t^{-1}$.

The *second* account of the notion of *difference* employs not an inverse for concatenation, but rather for union with the extension from sets to *multisets*, by which we mean sets in which each element is associated with a *multiplicity*, which we define to be a number in Z (and for the sake of brevity, we will typically refer to multisets as *m-sets*). This provision allows for an element to be a member of a multiset with negative multiplicity (or, much less interestingly, zero multiplicity). The multiplicity of a_i in A is indicated $\mu_A(a_i)$, where the subscript of μ can be omitted if context makes it clear which multiset we are concerned with. The union of two m-sets A and B is defined as the set union of the elements with non-zero multiplicity in either A or B, and the multiplicity of an element a in $A \cup B$ is defined as the sum of the multiplicities of a in A and in B. We adopt the convention that we have employed so far: the name of an individual string is expressed by a lower case letter, such as a or a_j, while the name of an m-set is expressed by a capital letter, such as A. It is both convenient and natural to indicate elements whose multiplicity is 1 with no special marking, and to indicate elements whose multiplicity is -1 with a prefixed minus sign: $-a$. A multiset $\{a, -b\}$ should *not* be thought of as containing an element denoted "$-b$"; rather, it contains an element b of multiplicity -1. This may lead to some confusion if we do not keep the definition clearly in mind. Thus it is true that $--A = A$, for any m-set A, but $\{--a\}$ does not equal $\{a\}$, and neither $--a$ nor $\{--a\}$ is a

2 There is no reason to be uncomfortable with the introduction of inverses with respect to both operations on the base set, as we do when we construct a field.

meaningful expression.[3]

We distinguish very sharply between the inverse for concatenation, expressed x^{-1}, and the inverse for multiset union, expressed with a leading minus sign, "$-x$". Let us say a few words about each approach before continuing

3 Concatenative inverses

In this approach, we augment the alphabet by adding for each letter in A an inverse. The inverse of a will be indicated a^{-1}, that of b as b^{-1} and so on: $aa^{-1} = a^{-1}a = \varepsilon$. The set of all inverses of the letters in A is noted as A^{-1}. We will use the capital script $\mathcal{A}$ to indicate the original alphabet A augmented with both the null symbol ε and the set of all inverses A^{-1}. $\mathcal{A}$ with the concatenation operator is thus isomorphic to the free group over the original alphabet A. The string cat has an inverse $(cat)^{-1}$, which equals $(t^{-1}a^{-1}c^{-1})$, and in similar fashion, $walking(ing)^{-1} = walk$. We can easily define a notion of difference of two strings, but we must distinguish between left difference and right difference. The *right difference* between *walks* and *walking* is $(walking)^{-1}walks = (ing)^{-1}s$, while the *left difference* between *walks* and *walking* is $walks(walking)^{-1}$, which does not simplify algebraically. The left difference between *view* and *preview* is $(pre)^{-1}$.

4 Multiset inverses

Multisets as we have defined them form a group under the operation of union, since any m-set has an inverse, in a natural way. If an m-set A consists of elements a_i, with corresponding multiplicities μ_i, then A's inverse is the m-set consisting of the same elements with multiplicities $-1 \cdot \mu_A$, and we will indicate the inverse of a multiset A as $-A$. We may also define $A - B$ as $A \cup -B$. For example, if $A = \{a, b, c\}$ and $B = \{a, b, d\}$, then $A - B = \{c, -d\}$. $B - A$, the difference between B and A, is $\{-c, d\}$.

3 The intersection operator from the algebra of sets does not carry over naturally to this version of multisets.

string s	string t	$\Delta_{\text{R}}(s,t) = \frac{s}{t}\text{R}$	$\Delta_{\text{L}}(s,t) = \frac{s}{t}\text{L}$
$walked$	$walking$	$\frac{ed}{ing}$	$\frac{walked}{walking}$
$walk$	$walking$	$\frac{\varepsilon}{ing}$	$\frac{walk}{walking}$
$walk$	$jump$	$\frac{walk}{jump}$	$\frac{walk}{jump}$
$walked$	$jumped$	$\frac{walked}{jumped}$	$\frac{walk}{jump}$
$remind$	$mind$	$\frac{remind}{mind}$	$\frac{re}{\varepsilon}$

FIGURE 1 – Some examples of string difference.

5 Pairwise differences

We return to our discussion of both approaches. In many of the cases that will be of interest, we are interested in the difference between two strings, or between two multisets, each consisting of just one string. The two approaches yield slightly different results for strings x and y. In the case of the concatenation inverse, we must specify whether we intend a left difference or the right difference of x and y; these are, respectively, xy^{-1} and $y^{-1}x$. We indicate these as $\Delta_{\text{L}}(x,y)$ and $\Delta_{\text{R}}(x,y)$. $\Delta_{\text{R}}(walks, walking) = \Delta_{\text{R}}(s, ing) = (ing)^{-1}s$. See figure 1 for further examples.

In the case of multiset union, the difference between the multiset $A = \{x\}$ and $B = \{y\}$ consists of a multiset with one element of multiplicity 1 and one of multiplicity -1: $\{x, -y\}$, and there is no distinction made between left and right difference. However, it is also true that the difference between *walks* and *walking* is $\{walks, -walking\} = \{walk\}\{s, -ing\}$, and the difference between *walking* and *walking* is $\{walk, -walk\}\{ing\}$.

It is convenient to express this more directly, and to use a notation that can be used for either approach to differences, and for this we use the notation $\frac{x}{y}$, $\frac{x}{y}\text{L}$, and $\frac{x}{y}\text{R}$. In the context of multiset inverses, we define $\{\frac{a}{b}\}$ as $\{a, -b\}$. In the context of concatenation inverse, we define $\frac{x}{y}\text{L}$ as $\Delta_{\text{L}}(x,y)$, and $\frac{x}{y}\text{R}$ as $\Delta_{\text{R}}(x,y)$.

We define the operation of concatenation of multisets of strings in the natural way, given a definition of concatenation of strings, in order

to ensure that concatenation distributes properly over the operation of (multiset) union. When it is helpful, we use the symbol $\times$ to mark concatenation, either of strings or of m-sets; when it leads to no confusion, we omit that symbol. The concatenation of two multisets of strings A and B is defined as the m-set of all strings of the form $a_i b_j$, $a_i \in A$, $b_j \in B$, and the multiplicity of $a_i b_j$ is $\mu_A(a_i) \cdot \mu_B(b_j)$. The following true statements illustrate the sort of descriptions we will explore:

1. $\{walks, jumps\} = \{walk, jump\}\{s\}$

2. $\{walk,\ jump\}\ \{s,\ ed,\ ing\} = \{walks,\ walked,\ walking,\ jumps,\ jumped, jumping\}$

3. $\{walks\} - \{jumps\} = \{walk, -jump\}\{s\} = \{\frac{walk}{jump}\}\{s\}$

4. $\{outperform\} - \{perform\} = \{out, -\varepsilon\}\{perform\} = \{\frac{out}{\varepsilon}\}\{perform\}$

5. $\{hard,\ soft\}\{\ en\ \}\{\ \varepsilon,\ s,\ ed,\ ing\ \} = \{harden,\ hardens,\ hardened,\ hardening,\ soften,\ softens,\ softened,\ softening\}$

6. $\{hard,\ harder,\ hardest,\ harden,\ hardens,\ hardened,\ hardening\} = \{hard\}\left\{\left\{\begin{array}{c}\varepsilon\\ er\\ est\end{array}\right\} \cup \{en\}\{\varepsilon, s,\ ed,\ ing\}\right\}$

7. $\{sing\} - \{sang\} = \{s\left\{\frac{i}{a}\right\} ng\}$

Observe that in the traditional semiring with concatenation of strings, union plays the role of addition and concatenation plays the multiplicative role, because concatenation is distributive over union, while union is not distributive over concatenation.[4] For a semiring to be a ring, there must be inverses for each element under the additive operation, which in our case reduces to the m-set union. Thus if we introduce m-set

4 That is, $\{a\} \cup (\{b\} \times \{c\}) = \{a, bc\}$ does not in general equal $(\{a\} \cup \{b\}) \times (\{a\} \cup \{c\}) = \{a, b\} \times \{a, c\} = \{aa, ac, ba, bc\}$.

inverses, the structure we are exploring is a ring, while if we adopt only concatenation inverses, we are not exploring a ring.

The examples in this paper are all taken from the standard orthography of English, but all of it applies equally to a transcription of a language in a phonetic or phonological form.

The ultimate goal of this work is to establish a counterpoint to generative grammar, in the following sense. Generative grammar takes as its task to elucidate the principles, both universal and language-particular, which accounts for the observed data on the basis of knowledge of the smallest relevant units in the language (typically morphemes and/or words). One analyzes a chosen utterance on the basis of the smallest underlying forms, which are already known to the linguist, employing the universal/language-particular principles that are called upon. This generative analysis, however, says nothing about the (epistemological) origin of these units and thereby leaves half of the account of the language untouched. American descriptivists, such as Zellig Harris, often referred to the tasks of segmentation and classification, with these tasks in mind, and the present paper aims to better understand what those principles would be when viewed from today's computational point of view.

6 Oppositions

6.1 The origins of the concept

The term *opposition* was proposed by Trubetzkoy in his major work, *Grundzüge der Phonologie* (Trubetzkoy, 1939), and it is closely related to the notion of difference between two elements in a language. Trubetzkoy proposed that two elements in a language can be put into a relationship with each other called an *opposition*, which consists of two things: first, a statement as to what they share in common, and secondly, a statement of what the first element possesses that the second does not, and of what the second possesses that the first does not, and the two of them together constitute a difference of the sort we have been discussing. If what we have called "properties" can be thought of as members of an m-set along the lines we described in the preceding section, then the statement of the properties that A possesses but that

B does not can be expressed as a multiset in which A's distinctive properties have a positive multiplicity and B's distinctive properties have a negative multiplicity. For Trubetzkoy and the structuralists who followed him, different items in the grammar of a language could be put into oppositions with one another: one phoneme could be put into opposition with another, one word with another, one case with another, and so on.

Consider a simple case, such as the phonological opposition $Opp(pat, mat)$ between the words pat and mat. What these two words have in common is $\_at$, which is to say, the sequence at positioned to the right of something else (the element of positioning is what is indicated by the underscore "_"), and their difference is the difference between a p and an m. What is the opposition $Opp(p, m)$ between p and m? That question gives rise to a second order opposition, that is, an opposition that arises because of the definition of a first order opposition; we shall see other sorts of second order oppositions below. The second order opposition here consists of a commonality (of p and m), and the differences. To describe this, we need recourse to features; features are the linguist's way of describing second order oppositions in phonology. What p and m have in common is a point of articulation – in particular, labial point of articulation. How they differ is that the first is voiceless and oral, and the second is nasal.

In this paper, we focus on oppositions between strings, and just stray briefly into the elements that permit us to discuss oppositions between individual elements, which are the "phonological features" that grew out of Trubetzkoy's and Jakobson's conception of phonology.

We may return now to this question: what *is* an opposition between two strings? By definition it is two things: a statement of what the pair of words have in common, and a statement of how each differs from the other. In the case of strings, a natural statement (but not the only reasonable statement) of what they have in common is to produce, first, a *substring* present at the left edge of both strings, or at the right edge of both strings, along with a statement as to where that common substring appears in the two words; and second, a statement of how they *differ*, which is a string difference of what remains if we remove the

string they have in common.[5] In the case we are considering, there is a natural connection between this and the distributivity of concatenation over m-set union, since extracting a maximal common substring at an edge is no different from using the distributive identity in a maximal fashion; that is, just as $\{walked, walking\} = \{walk\}\{ed, ing\}$, so $\{walked\} - \{walking\} = \{\frac{walked}{walking}\} = \{walk\}\{\frac{ed}{ing}\}$.

When we go beyond phonology, we compare more than simply strings or structures of sounds; we compare words with word-particular information (which are called **semantic** and **morphological** much of the time), since even the information that we consider syntactic we usually refer to as *morphosyntactic* when we consider words out of syntactic context.

When we consider the opposition $(walks/walking)$, we ask what they have in common and what they do not have in common. What they have in common is a string, $walk$, which precedes their difference, and semantic information, as well as the morphosyntactic information that we call "Verb". They differ with the string opposition (s/ing) and morphosyntactic features that differ between these two suffixes.

In this case, we do not get much in return for asking what (s/ing) have in common and how they differ (this is different from the phonological case, where there is a difference between information in a string and information "inside" a phoneme, as we noted when asking what the difference was between p and m.) In this case, if the two suffixes have some morphosyntactic information in common, it is natural to associate it with the stem, i.e., the commonality.

As Trubetzkoy notes, there are cases where there is a natural ordering of the two elements in the difference. In some cases, the opposition-difference is something versus nothing $(walking/walk)$ and in some cases the opposition difference is "less" versus "more". This latter kind of opposition is natural and even central in semantics, and more limited in the domain of phonology.

Thus we return to the central example of a binary word-based

5 Lee (2002) provides a detailed examination of alternative definitions of string commonality in this context. An important point to bear in mind is that while asking what two objects have in common is a perfectly meaningful question, it may have more than one reasonable answer.

opposition as a common stem plus an ordered pair of two affixes, specified either as prefixal or as suffixal (i.e., indicating if they precede or follow the common stem). An opposition is an ordered pair, then, of two things: a commonality and a difference, and so from a logical point of view an opposition is an ordered pair of a commonality and a difference. This definition is intended to be extremely general, and by no means restricted to strings. Given its importance here, we propose to indicate this with its own notation: in an opposition between X and Y, indicated $[A, \frac{B}{C}]_{Opp}$, A is what X and Y have in common, B is what X has and Y does not, and C designates what Y has and X does not. That statement summarizes the result of this section.

Before continuing, let us review some classic observations about morphological (i.e., word-internal) structure. The set of words $\{truck, train, travel, trip\}$ equals $\{tr\}\{uck, ain, avel, ip\}$ (and, as one can easily see, all of the words share a semantic component somehow relating to locomotion, to put it awkwardly). But this analysis is amusing rather than insightful (or grammatical). There are two ways of expressing why this analysis is not significant. In the terms that we are proposing, it is because the commonalities of any pair of these four words is different (the opposition of $truck$ and $train$ begins with the commonality of the two words, which includes the shared meaning of "vehicle", while the opposition of $truck$ and $trip$ includes nothing of that sort in the commonality). A related but nonetheless distinct way of expressing the irrelevance lies in Greenberg's condition on word analysis (morphemic analysis), which requires that a morphemic analysis contains minimally two items, as expressed in the classic Greenberg rectangle (see Greenberg, 1960).

$walk$	ed
$jump$	ing

which could also be expressed as $\left\{ \begin{array}{c} walk \\ jump \end{array} \right\} \left\{ \begin{array}{c} ed \\ ing \end{array} \right\}$.

We believe that morphology emerges with the presence of at least two oppositions which share the same difference. It is natural to define

an operation of union on oppositions, but only on pairs (or sets) of oppositions that have identical differences, such as $[W, \frac{Y}{Z}]_{Opp}$ and $[X, \frac{Y}{Z}]_{Opp}$. Here we define their union as:

$$[W, \frac{Y}{Z}]_{Opp} \cup [X, \frac{Y}{Z}]_{Opp} := [W \cup X, \frac{Y}{Z}]_{Opp}$$

Given the two oppositions $[walk\_, \frac{ed}{ing}]_{Opp}$ and $[jump\_, \frac{ed}{ing}]_{Opp}$, their union is $[walk\_ \cup jump\_, \frac{ed}{ing}]_{Opp}$. With this, we turn to the notion of a paradigm in a morphology, where that property of opposition union plays a central role.

6.2 Oppositions within a paradigm

Let us begin by defining a paradigm as simply a set of words, recognizing that in the real world a good deal more is intended when one speaks of a paradigm. Two typical paradigms that we will be interested in is $P_{walk} = \{walk, walks, walked, walking\}$ and $P_{move} = \{move, moves, moved, moving\}$.

Let us first consider the opposition of a paradigm with itself, a *self-opposition*, which can be naturally thought of as an array of all of the pairwise oppositions of distinct members of the paradigm, as in figure 2 (where we have stacked the arguments for ease of presentation), a sort of outer product of the paradigm with itself, where we use Φ to indicate the opposition of something with itself, an essentially useless object.

P_{walk}	$walk$	$walks$	$walked$	$walking$
$walk$	Φ	$Opp\begin{pmatrix} walk, \\ walks \end{pmatrix}$	$Opp\begin{pmatrix} walk, \\ walked \end{pmatrix}$	$Opp\begin{pmatrix} walk, \\ walking \end{pmatrix}$
$walks$	$Opp\begin{pmatrix} walks, \\ walk \end{pmatrix}$	Φ	$Opp\begin{pmatrix} walks, \\ walked \end{pmatrix}$	$Opp\begin{pmatrix} walks, \\ walking \end{pmatrix}$
$walked$	$Opp\begin{pmatrix} walked, \\ walk \end{pmatrix}$	$Opp\begin{pmatrix} walked, \\ walks \end{pmatrix}$	Φ	$Opp\begin{pmatrix} walked, \\ walking \end{pmatrix}$
$walking$	$Opp\begin{pmatrix} walking, \\ walk \end{pmatrix}$	$Opp\begin{pmatrix} walking, \\ walks \end{pmatrix}$	$Opp\begin{pmatrix} walking, \\ walked \end{pmatrix}$	Φ

FIGURE 2 – Self-opposition of a paradigm.

In looking at an array of pairwise oppositions, it is natural to separate it into two arrays, one for the commonalities, and one for the differences, and we have done just this, showing the commonalities in figure 3, and the differences in figure 4 (we omit diagonal elements throughout). (Here as elsewhere, the expression $\frac{a}{b}$ denotes $\{a, -b\}$, in the multiset interpretation, or $\{a, b^{-1}\}$ in the concatenation interpretation.)

P_{walk}	$walk$	$walks$	$walked$	$walking$
$walk$	Φ	$walk\_$	$walk\_$	$walk\_$
$walks$	$walk\_$	Φ	$walk\_$	$walk\_$
$walked$	$walk\_$	$walk\_$	Φ	$walk\_$
$walking$	$walk\_$	$walk\_$	$walk\_$	Φ

FIGURE 3 – Left-edge commonalities in the paradigm P_{walk}.

P_{walk}	$walk$	$walks$	$walked$	$walking$	
$walk$	Φ	$\frac{\varepsilon}{s}$	$\frac{\varepsilon}{ed}$	$\frac{\varepsilon}{ing}$	$\boldsymbol{\varepsilon}$
$walks$	$\frac{s}{\varepsilon}$	Φ	$\frac{s}{ed}$	$\frac{s}{ing}$	$\boldsymbol{s}$
$walked$	$\frac{ed}{\varepsilon}$	$\frac{ed}{s}$	Φ	$\frac{ed}{ing}$	$\boldsymbol{ed}$
$walking$	$\frac{ing}{\varepsilon}$	$\frac{ing}{s}$	$\frac{ing}{ed}$	Φ	$\boldsymbol{ing}$
	$\boldsymbol{\varepsilon}$	$\boldsymbol{s}$	$\boldsymbol{ed}$	$\boldsymbol{ing}$	

FIGURE 4 – Right differences in P_{walk}.

Thus prefixes or suffixes emerge from the description of the differences between members of a paradigm in an outer product of the oppositions of the members of the paradigm.

An opposition is by its nature a relation between two objects, but a paradigm is in general a larger set of forms (larger than two, that is), and part of what holds it together conceptually is that all of its members share something in common – which we often call its stem. It is thus natural to expand the concept of a binary opposition to a paradigm-like

set of forms under the condition that each pair of items is analyzed as an opposition, but one in which the all pairs share the same commonality (here, the stem). We will refer to such cases, the sort in figures 3 and 4, as *pure paradigms.*

e-final verbal pattern		
$move$	$love$	$hate$
$moves$	$loves$	$hates$
$moved$	$loved$	$hated$
$moving$	$loving$	$hating$

FIGURE 5 – English e-final verb stems.

In the case of verbs in English such as $walk$ or $jump$, the array of commonalities is constant throughout, but in the case of other verbal paradigms, the commonalities in some pairs is different from the commonalities in other pairs. We consider first e-final stems, as illustrated in figure 5, and analyzed in figure 6, and we see that in some cases, the commonality is mov and in others it is $move$. From a simple logical point of view, satisfying the condition that all commonalities be the same appears to be met by making the stem smaller and smaller, so to speak: in this case, making it mov, and changing the analysis to figure 7.

In work not reported here, we extend the computation of oppositions to the case of opposition between two self-oppositions of paradigms; that is, in the case illustrated here, we define the opposition between (for example) P_{move} and P_{walk}.

7 Conclusion

We hope to have provided a small view of a mathematical way of understanding differences that arise in a systematic way in natural language. In work in progress, we extend this notion of difference from differences within paradigms to differences across paradigms, to better understand how languages employ large families of pairs of differences, where the oppositions within the families are constant, and where the

P_{move}	*move*	*moves*	*moved*	*moving*
move		*move*	*move*	*mov*
moves	*move*		*move*	*mov*
moved	*move*	*move*		*mov*
moving	*mov*	*mov*	*mov*	

P_{move}	*move*	*moves*	*moved*	*moving*	
move	Φ	$\frac{\varepsilon}{s}$	$\frac{\varepsilon}{d}$	$\frac{e}{ing}$	$\boldsymbol{e, \varepsilon}$
moves	$\frac{s}{\varepsilon}$	Φ	$\frac{s}{d}$	$\frac{es}{ing}$	$\boldsymbol{s, es}$
moved	$\frac{d}{\varepsilon}$	$\frac{d}{s}$	Φ	$\frac{ed}{ing}$	$\boldsymbol{d, ed}$
moving	$\frac{ing}{e}$	$\frac{ing}{es}$	$\frac{ing}{ed}$	Φ	$\boldsymbol{ing}$
	$\boldsymbol{e, \varepsilon}$	$\boldsymbol{e, es}$	$\boldsymbol{d, ed}$	$\boldsymbol{ing}$	

FIGURE 6 – Analysis 1 of an *e*-final stem: *not* a pure paradigm (mixed stems).

M_{move}	*move*	*moves*	*moved*	*moving*
move		*mov*	*mov*	*mov*
moves	*mov*		*mov*	*mov*
moved	*mov*	*mov*		*mov*
moving	*mov*	*mov*	*mov*	

M_{move}	*move*	*moves*	*moved*	*moving*	
move	Φ	$\frac{e}{es}$	$\frac{e}{ed}$	$\frac{e}{ing}$	$\boldsymbol{e}$
moves	$\frac{es}{e}$	Φ	$\frac{es}{ed}$	$\frac{es}{ing}$	$\boldsymbol{es}$
moved	$\frac{ed}{e}$	$\frac{ed}{es}$	Φ	$\frac{ed}{ing}$	$\boldsymbol{ed}$
moving	$\frac{ing}{e}$	$\frac{ing}{es}$	$\frac{ing}{ed}$	Φ	$\boldsymbol{ing}$
	$\boldsymbol{e}$	$\boldsymbol{es}$	$\boldsymbol{ed}$	$\boldsymbol{ing}$	

FIGURE 7 – Analysis 2 of *e*-final stem, with stem *mov*-, a pure paradigm.

cross-family differences are themselves bounded in certain respects. For example, there are two distinct families of inflections for French verbs including *choisir* and *partir*, respectively. Each family is defined by the pairwise differences within it, but a higher order set of differences can be computed that deals with the differences across the two families. We have employed some of this work in software for learning morphology, employing Minimum Description Length methods for calculating the information content of oppositions and sets of oppositions.

Acknowledgements

Thanks to Paul Goldsmith-Pinkham and Jason Riggle for helpful comments on an early version of this paper.

References

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26(3):178–194.

Lee, J. L. (2002). *Morphological paradigms: Computational structure and unsupervised learning*. PhD thesis, University of Chicago.

Trubetzkoy, N. (1939). *Grundzüge der Phonologie*. Number 7 in Travaux du Cercle Linguistique de Prague. Cercle linguistique de Prague, Prague.

Cahiers du CLSL, n° 69, 2025, pp. 73-96

A framework for spatial clustering of textual objects: applications in topic clustering and text segmentation

Guillaume Guex
University of Lausanne
guillaume.guex@unil.ch

Abstract

We present a general, classical, framework of spatial clustering which can be applied to various textual objects (e.g. character n-grams, words, sentences). This framework proposes to cluster objects according to users defined linguistic similarity, while keeping a spatial coherence of objects among clusters. Two methods are derived from this formalism: *SpatialWord*, which applies to word-tokens, and *SpatialSent*, operating on sentences, which both balance between semantic similarities of objects and their position along the textual sequence. We show that these unsupervised methods, along with semi-supervised variants, can perform jointly two operations often achieved individually by methods in literature: (1) the extraction of a desirable number of topics from a document along with list of words to interpret them; and (2) the textual segmentation of the document reflecting these extracted topics. Case studies show that these methods perform competitively against state-of-the-art methods on baseline datasets.

1 Introduction

Automatically revealing topics in a document has been of great value for domains such as information retrieval, question answering or digital humanities, as it can effectively extract information from a document without actually reading it (distant reading). Historically, topic modeling approaches, such as *Latent Dirichlet Allocation (LDA)* (Blei et al., 2003) or *Non-negative Matrix Factorization (NMF)* (Arora et al., 2012), considered documents as bags-of-words, and supposed that similar

topics are found in documents having comparable vocabulary usages. These approaches give affinity weights (e.g. probabilities) to words depending on the topic, which allow its identification, but without taking into account one of the main arrangements made by authors in their text: topics are generally found in coherent, contiguous passages. Retrieving passages addressing a particular extracted topic is generally tedious with such approaches, as contiguous words can belong to very different topics. By contrast, text segmentation methods (Choi, 2000; Eisenstein & Barzilay, 2008; Glavaš et al., 2016; Koshorek et al., 2018; Riedl & Biemann, 2012), which are also used to automatically extract information from documents, mainly use the structure of the text, i.e. the relative position of textual elements (tokens, sentences, paragraphs) in the sequence. They generally use some detection of semantic shift to place breakpoints in documents, resulting in a segmentation that can reflect the topical structure intended by the author. However, unlike topic modeling approaches, these methods generally do not label the resulting segments, and are unable to quickly summarize the main topics found in a document with lists of most used words. While methods combining approaches, i.e. finding a textual segmentation and assigning labels to segments, are obviously valuable, they are rare to find. Some works use segmentation with a combined text classification (Agarwal & Yu, 2009; Arnold et al., 2019; Chen et al., 2009; Tepper et al., 2012), using shared knowledge over the whole corpus, but none, to our knowledge, use an unsupervised approach on a single file.

We propose here to use a very general and flexible framework based on *spatial autocorrelation*, originally used in Bavaud et al. (2015) and Ceré & Bavaud (2018) and previously inspired from Anselin (2010) and Cressie (1993), which allows *spatial clustering* methods of various textual objects, along with *semi-supervised classification* variants. Methods derived from this formalism are able to extract topics on a single document, and can be tuned to force a spatial coherence of these topics in the text, hence finding segments of text covering each topic. They can be applied to different textual objects: character n-grams, word-tokens, sentences, paragraphs; as long as two quantities are defined on them: (1) a *similarity* (or dissimilarity) between elements,

which can reflect semiotic, phonological, or semantic affinities between textual objects; (2) a *proximity structure*, defining how elements relate to each others in the textual sequence. The proposed framework applied on n objects with m groups given will result in a $(n \times m)$ fuzzy membership matrix, noted $\mathbf{Z} = (z_{ig})$, verifying $z_{ig} \geq 0$ and $\sum_g z_{ig} = 1$, where z_{ig} reflects the membership percentage of object i in group g. Along with the number of desired groups m, methods have two main hyperparameters, α and β, whose tuning can balance, on one hand, between the importance of the similarity of items vs their proximity, and, on the other hand, between the fuzziness vs the crispiness of group memberships. In case studies, we illustrate two methods derived from this formalism: a *semantic clustering of word-tokens* and a *semantic clustering of sentences* in a given document, along with their *semi-supervised classification variants*. We show that these methods can be used on *topic clustering* and *text segmentation* tasks, extracting interpretable topics along with text segments that cover them. For these tasks, our methods are compared with cutting-edge methods on gold standard datasets, showing that, while not state-of-the-art, they can compete against the best methods. Section 2 explains the framework used by our methods, section 3 explores case studies, and section 4 draws general conclusions. All datasets, Python scripts and results can be found in the Github repository of the article.[1]

2 Formalism

2.1 General framework

2.1.1 Dissimilarity and exchange matrices

While this article focuses on semantically clustering *word-tokens* or *sentences*, the framework used here can be defined in very general terms, as found in, e.g., Bavaud et al. (2015) and Ceré & Bavaud (2018). Consider n *objects*, indexed by $i \in \{1, \ldots, n\}$, with their *vector of relative weights* $\mathbf{f} = (f_i)$, where $f_i > 0$ and $\sum_i f_i = 1$, and the following matrices:

– An $n \times n$ symmetric *squared Euclidean dissimilarity matrix*

1 https://github.com/gguex/SemSim_AutoCor.

$\mathbf{D} = (d_{ij})$, verifying $d_{ij} \geq 0$, containing pairwise dissimilarities between these objects.

– An $n \times n$ symmetric matrix of joint probabilities $\mathbf{E} = (e_{ij})$, called *exchange matrix*, verifying $e_{ij} \geq 0$ and $e_{i\bullet} = e_{\bullet i} = f_i$ ("$\bullet$" refers to a sum over the replaced index), containing spatial relationships between objects. Sometimes, we use the associated *Markov chain transition matrix* $\mathbf{W} = (w_{ij})$, defined with $w_{ij} := e_{ij}/f_i$. The margins of this matrix must contain object weights in order for the functionals to be formally defined.

2.1.2 Membership matrix and functionals

A *fuzzy clustering* of these n objects into m groups can be defined by a $n \times m$ *membership matrix* $\mathbf{Z} = (z_{ig})$, with $z_{ig} \geq 0$ and $z_{i\bullet} = 1$, whose components represent the membership of object i to group g. The membership matrix defines the *relative group weights vector* $\boldsymbol{\rho} = (\rho_g)$, with $\rho_g := \sum_i f_i z_{ig}$ and m vectors of *within-group distribution* $\mathbf{f}^g = (f_i^g)$ with $f_i^g := f_i z_{ig}/\rho_g$. Different functionals can be computed from a membership matrix $\mathbf{Z}$.

The *within-group inertia* is defined as

$$\Delta_W[\mathbf{Z}] := \sum_g \rho_g \Delta_g \quad \text{where} \quad \Delta_g = \frac{1}{2} \sum_{ij} f_i^g f_j^g d_{ij} \ . \tag{1}$$

A low within-group inertia reflects homogeneity between objects of the same group, as defined by the dissimilarity matrix $\mathbf{D} = (d_{ij})$.

The *generalized cut* reads

$$\mathcal{C}^\kappa[\mathbf{Z}] := \sum_g \frac{\rho_g^2 - e(g,g)}{\rho_g^\kappa} \quad \text{where} \quad e(g,g) := \sum_{ij} e_{ij} z_{ig} z_{jg} \ . \tag{2}$$

A low generalized cut functional indicates strong neighborhood relationships between tokens of the same group, as defined by the exchange matrix $\mathbf{E} = (e_{ij})$. The hyperparameter $\kappa \in [0, 1]$ allows

us to interpolate objects the *N-cut objective* (Shi & Malik, 2000) when $\kappa = 1$ and the *modularity criterium* (Newman, 2006) when $\kappa = 0$.

The token-group dependency can be expressed by the *mutual information*

$$\mathcal{K}[\mathbf{Z}] := \sum_{ig} \rho_g f_i^g \log\left(\frac{f_i^g}{f_i}\right) = \sum_{ig} f_i z_{ig} \log\left(\frac{z_{ig}}{\rho_g}\right) \tag{3}$$

which is low if distributions f_i^g correspond to f^i, i.e. f_i^g are independent of group g. Therefore, low mutual information indicates fuzziness in group memberships.

Finally, by combining all previous functionals, we can define the *free energy* with

$$\mathcal{F}[\mathbf{Z}] := \beta\Delta_W[\mathbf{Z}] + \frac{\alpha}{2}\mathcal{C}^\kappa[\mathbf{Z}] + \mathcal{K}[\mathbf{Z}] \ . \tag{4}$$

Searching the membership matrix $\mathbf{Z}$ minimizing this functional results in a fuzzy clustering of objects depending on hyperparameters α, β and κ. An interpretation of hyperparameter effects in the case of textual data can be found in section 2.2.4.

2.1.3 Finding a local minima

Canceling the derivative of the free energy with respect to z_{ig} under the constraints $z_{i\bullet}$ yields the minimization condition

$$z_{ig} = \frac{\rho_g \exp(-h_{ig})}{\sum_h \rho_h \exp(-h_{ih})} \tag{5}$$

where

$$\rho_g[\mathbf{Z}] := \sum_i f_i z_{ig} \tag{6}$$

$$h_{ig}[\mathbf{Z}] := \beta d_i^g + \alpha\rho_g^{-\kappa}(\rho_g - \sum_j w_{ij} z_{jg}) - \frac{\alpha\kappa}{2}\rho_g^{-\kappa-1}\left(\rho_g^2 - e(g,g)\right) \tag{7}$$

with $d_i^g = \sum_j f_j^g d_{ij} - \Delta_g$ the squared Euclidean dissimilarity from i to the centroid of group g. These Equations define an iterative procedure converging to a local minimum for $\mathcal{F}[\mathbf{Z}]$: a random membership matrix is taken as $\mathbf{Z}^0$, ρ_g^t and h_{ig}^t are computed with (6) and (7) respectively, and $\mathbf{Z}^{t+1}$ is given by (5).[2] Pseudo-code for the algorithm can be found in appendix A.1.

2.1.4 Semi-supervised framework

Working with the membership matrix $\mathbf{Z}$ for objects allows us to easily adapt the algorithm in a semi-supervised framework. Let $\mathcal{T}$ be the group of *tagged objects*, which consists in m disjoint subgroups $\mathcal{T} = \cup_{g=1}^m \mathcal{T}_g$, where $\mathcal{T}_g$ contains objects which should be in group g. The initial membership values z_{ig}^0 for tagged object $i \in \mathcal{T}$ are then set to :

$$z_{ig}^0 = \begin{cases} 1 & \text{if } i \in \mathcal{T}_g \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Moreover, at the end of each iteration, the tagged objects are reset to their initial values z_{ig}^0, forcing them to be in their respective group. If the generalized cut functional is high enough (i.e. α in (4) is large), these tagged objects should act as anchor points, "spreading" their labels to their neighbors.

2.1.5 Complexity and scaling

The algorithm complexity is $\mathcal{O}(n^2 m)$, which can be problematic in case of large datasets. However, this issue can be alleviated by decomposing the dataset into small overlapping blocks, and running the algorithm on each of them independently, while transferring the labels from one block to another by fixing membership of objects at their intersection. Formally, our dataset $\mathcal{D}$ can be decomposed into p blocks $\mathcal{B}_k$ of size $n_b < n$, with $\mathcal{D} = \cup_{k=1}^p \mathcal{B}_k$ and $\mathcal{B}_k \cap \mathcal{B}_{k-1} \neq \emptyset,\ \forall k \in \{2, \ldots, p\}$. We first find the memberships $z_{ig}^{\mathcal{B}_1}$ of the objects in the group $\mathcal{B}_1$, and

2 During applications, $\mathbf{Z}^0$ is picked with $\mathbf{f}^g$ close the uniform distribution for every g, and the new membership matrix is computed with $\lambda \mathbf{Z}^{t+1} + (1-\lambda)\mathbf{Z}^t$ with $\lambda \in (0, 1]$ a decreasing adaptative learning parameter which allows the algorithm to reach the bottom of "valleys" of high gradients.

proceed sequentially by fixing $z_{ig}^{\mathcal{B}_k} = z_{ig}^{\mathcal{B}_{k-1}},\ \forall i \in \mathcal{B}_k \cap \mathcal{B}_{k-1}$. The algorithm is then $\mathcal{O}(n_b^2 pm)$, but some performance is lost in the process. We examine the speed/performance trade-off of this procedure in Section 3.4.3.

2.2 Textual data

The framework of section 2.1 can be adapted to textual data, where objects can be, e.g., character n-grams, word-tokens, sentences, or paragraphs, as long as we can define: (1) dissimilarities reflecting linguistic differences between items; and (2) a spatial proximity structure, indicating how objects interact with each other in the textual sequence. The relative weights f_i for textual objects can be defined as uniform or proportional to their length (e.g. the number of words in a sentence).

2.2.1 Semantic (dis)similarities

When working with textual objects, the matrix $\mathbf{D}$ should represent *linguistic dissimilarities* between objects, which can be defined notably at the semiotic, phonological, or semantic level, depending on objects and applications. We can generally construct these dissimilarities from external sources, giving either dissimilarities or similarities. Dissimilarities d_{ij} can be obtained from similarities s_{ij} with, e.g., $d_{ij} = \max_{kl} s_{kl} - s_{ij}$. In this article, we exclusively work with word-tokens and sentences, and with dissimilarities defined at the semantic level. For tokens, semantic similarities can be constructed with *WordNet* (Fellbaum, 1998), *type-based Word Embeddings* (Bojanowski et al., 2017; Mikolov et al., 2013a; Pennington et al., 2014), or *Transformers* (Devlin et al., 2019). Note that when using type-based Word Embeddings, these dissimilarities are defined between *types*, and every pair of tokens having the same signature of types will yield the same dissimilarity. If we consider sentences, there are also many models permitting to build semantic similarities, such as *Sentence Embeddings* (Mikolov et al., 2013b; Reimers & Gurevych, 2019).

2.2.2 Spatial structure

Note that the structure of textual objects is a sequence that can be indexed by $i \in \{1, \ldots, n\}$. The spatial structure, encoded by the ex-

change matrix $\mathbf{E}$, should reflect how these objects relate to each other in the textual sequence. It could be crafted carefully, by using, e.g. a syntactic dependency parser in order to weight token relationships accordingly. However, we chose to construct it uniquely on the neighborhood relationships as it already shows satisfying results. Note that because of the constraint $e_{i\bullet} = e_{\bullet i} = f_i$ and because of boundary effects, the computation of these matrices is not trivial, even for uniform weights. Our suggestion is to compute textual exchange matrices as proposed in Ceré & Bavaud (2018), also seen in A.2.

1. The *uniform* exchange matrix $\mathbf{E}^{\mathrm{U}(r)}$ with range $r \in \mathbb{N}$, where $e_{ij}^{\mathrm{U}(r)}$ is constant for $j \in [i-r, i+r]$ and 0 otherwise. This matrix can be computed with the Metropolis-Hastings algorithm.

2. The *diffusive* exchange matrix $\mathbf{E}^{\mathrm{D}(t)}$, closely related to the diffusion map kernel (Nadler et al., 2005), with diffusive time factor $t > 0$. For a large enough t, the strength of links between i and its neighbors j will be normally shaped, thus exponentially decreasing with $|i-j|$.

The choice between $\mathbf{E}^{\mathrm{U}(r)}$ and $\mathbf{E}^{\mathrm{D}(t)}$, as well as their ranges r or t, can be considered as hyperparameters to be tuned. Computations for these matrices are given in Appendix A.2.

2.2.3 Topic interpretation

When semantic dissimilarities are used, the resulting clusters can be interpreted as *topics*. When using word-tokens as objects, the membership matrix $\mathbf{Z}$ has the following interpretation :

$$z_{ig} = \mathrm{P}(T_i = g) \tag{9}$$

where T_i is the variable containing the topic of token i. Note that with our algorithm, a token can be a part of multiple topics (fuzziness) and also depends on its neighborhood (spatial component), reflecting the fact that generally, a particular topic covers several contiguous textual objects.

When using a topic modeling approach, in order to interpret topics, we

are interested in the probability of each word-type constituting topics or, conversely, the probability of being in a certain topic when using a specific word-type. In other words, we need to compute:

$$\mathrm{P}(T = g|W = w) = \text{“Probability to be in topic } g \text{ when the type is } w\text{”} \tag{10}$$

$$\mathrm{P}(W = w|T = g) = \text{“Probability to draw type } w \text{ when the topic is } g\text{”} \tag{11}$$

where T is the variable containing the topic and W the variable containing the word-type for a randomly drawn token in the text. It is possible to express these quantities with the components of the matrix $\mathbf{Z}$ and using the variable W_i, containing the word-type at position i:

$$\mathrm{P}(T = g|W = w) = \frac{1}{n_w} \sum_{i|W_i=w} \mathrm{P}(T_i = g) = \frac{1}{n_w} \sum_{i|W_i=w} z_{ig} \tag{12}$$

$$\begin{aligned} \mathrm{P}(W = w|T = g) &= \frac{\mathrm{P}(T = g|W = w) \cdot \mathrm{P}(W = w)}{\mathrm{P}(T = g)} \\ &= \frac{\frac{1}{n_w} \sum_{i|W_i=w} \mathrm{P}(T_i = g) \cdot \frac{n_w}{n}}{\frac{1}{n} \sum_i \mathrm{P}(T_i = g)} = \frac{\sum_{i|W_i=w} z_{ig}}{z_{\bullet g}} \end{aligned} \tag{13}$$

where n_w designates the number of occurrences of word-type w. In case studies, section 3.4.1, we see that Equation (12) helps to reflect the vocabulary specificity of each topic, whereas Equation (13) gives clues about the writing style of each topic, which can remain similar among topics in some cases.

2.2.4 Hyperparameters effects

The result of the clustering depends on hyperparameters α, β and κ in (4). Increasing α relatively to β favors groups containing long sequences of textual objects, while increasing β for a fixed α strengthens linguistic homogeneity inside clusters. The limit $\beta \to \infty$ corresponds to the case of a *K-means* clustering using only the linguistic dissimilarities defined on textual objects. Decreasing both α and β increases the

fuzziness of groups, resulting in a mixture of groups defined over textual objects. Figure 1 gives insights on how the clustering behaves depending on α and β. The hyperparameter κ controls the spatial objective by interpolating between the N-cut objective and the modularity criterion, and its effect is more difficult to interpret. These three hyperparameters are tuned with grid search in case studies.

FIGURE 1 – Semantic clustering of word-tokens into 3 groups on a part of a MANIFESTO file (see section 3). Left: β is high relatively to α, resulting in small sequences. Middle: α is low relatively to β resulting in large sequences. Right: α and β are low, resulting in high group fuzziness, represented by the mixture of colors.

3 Case studies

3.1 Tasks

In case studies, in order to compare methods derived from our formalism with existing ones, we are interested in how they perform in two particular tasks, namely *topic clustering* and *text segmentation*.

Topic clustering, to our knowledge, is not a term used in literature, although existing methods, such as *LDA* (Blei et al., 2003) or *NMF* (Arora et al., 2012), can be adapted to perform it. In this article, we define it as the task of assigning groups to textual objects, with unsupervised methods, so that objects in the same group express a similar topic. Unlike topic modeling approaches, an affinity matrix is computed between every word-token and topic, taking into consideration the wor-token's position in text, and not only between word-types and topics. Validation of this task must use unsupervised measures of adequation, such as the *Normalized Mutual Information (NMI)* (Strehl & Ghosh, 2002) used

here. Note that measures of self-coherence, such as the *Perplexity* often used in topic modeling, do not apply here as our model is not generative. Section 3.4.1 explores the topic clustering capabilities of our methods. By contrast, the text segmentation task is well known in literature (Arnold et al., 2019; Chen et al., 2009; Choi, 2000; Eisenstein & Barzilay, 2008; Glavaš et al., 2016; Koshorek et al., 2018; Riedl & Biemann, 2012). It generally consists in finding breakpoints in a text, such that the resulting segments address different topics. Supervised and unsupervised methods for this task exist. Generally, apart from a few exceptions (e.g. Arnold et al., 2019; Chen et al., 2009), these methods do not label segments with a topic. Hence, the result does not indicate if the document consists of two alternating topics, a number of topics equal to the number of segments, or a situation in-between. Validation is generally made through the use of the P_k index (Beeferman et al., 1999) or the *Window diff (WD)* index (Pevzner & Hearst, 2002), measuring if separation marks between segments correspond to ground truth. Section 3.4.2 will compare the text segmentation performances of our algorithm against cutting edge methods.

To our knowledge, no other algorithm performs both tasks at the same time in an unsupervised way, especially *without using contrastive information between documents*. While sharing information over several documents can be an asset in some cases, and usually give better performances, using our method enables the user to find text segments, with their assigned topics (with word-types defining them, see section 2.2.3), in a single document without any training (though the choice of hyperparameters can still be subject to tuning).

3.2 Corpora

While it is possible to create artificial datasets to validate both topic clustering and text segmentation tasks (Choi, 2000), we favor here real datasets, as methods seem to give over-confident results on artificial ones (Glavaš et al., 2016). We will use five datasets, which can all be found in our Github repository.

The MANIFESTO dataset (Volkens et al., 2017)[3] consists in textual

3 https://manifesto-project.wzb.eu/, accessed Jan. 2025.

parties policy positions from different countries. Some of them are manually annotated with topics along the text, which is a premium resource to test methods in a topic clustering or classification task. Topical annotations are divided between 7 super-topics and several sub-topics. We chose here to use only super-topic classes, as the number of groups seems reasonable for our algorithm. For coherence in language and culture, we extracted the annotated documents from US parties, which corresponds to 9 documents: the manifestos from the Democratic Party from 1992, 2004, 2012, 2016, and 2020, and the Republican Party from 2004, 2008, 2012, and 2016.

The WIKI50 dataset, introduced in (Koshorek et al., 2018), consists in a set of 50 randomly sampled test documents from the largest WIKI-727K. It consists in 50 English Wikipedia articles and their segmentation corresponding to their table of content.

The CITIES and ELEMENTS datasets, introduced in Chen et al. (2009), are two datasets which are also extracted from English Wikipedia. The first consists in 100 articles about large cities, and the second in 119 articles about chemical elements in the periodic table.

Finally, the CLINICAL dataset, introduced in Eisenstein & Barzilay (2008), consists in 226 chapters extracted from the *Clinical Textbook*, which are mainly used in the topic segmentation task because the different sections are not labeled.

All these datasets are well designed for text segmentation, but, with the exception of the MANIFESTO dataset, they are less suitable for topic clustering, especially if information is not shared across documents. As a matter of fact, each document extracted from Wikipedia (WIKI50, CITIES, ELEMENTS), when looked at individually, has a unique label for each segment. This means that, when a method of clustering is applied on a unique file, these documents operate like the CLINICAL dataset with unlabeled segments: the number of groups found in a document is always equal to the number of segments. This situation is not ideal, as a topic should be a recurring subject, appearing at multiple places in a document, as found in the MANIFESTO dataset. Nevertheless, because of the lack of other real datasets and because these corpora are largely used to evaluate methods in literature, we will use them here

for comparison purposes.
All datasets are preprocessed the same way: case is lowered, stop-words, numbers and punctuation marks are removed, while information about where each sentence ends is kept in order to apply the method on sentences.

3.3 Methodology

For both tasks, we used two versions of our algorithm, which work on two different textual objects: we named *SpatialWord* our method when applied on word-tokens, and *SpatialSent* when used on sentences. Relative weight $\mathbf{f}$ is defined as uniform for word-tokens, and proportional to the number of words for sentences. For each document, the real number of groups is given to our method. Both methods are tried with a semi-supervised version with 5% and 10% random labeling rate.
For word-tokens, we wanted to use semantic similarities which do not use their local context, as we wanted to rely solely on the exchange matrix to express spatial dependencies. This excludes word-token embeddings, e.g. based on BERT (Devlin et al., 2019), because of their use of the context of a token to build its vector. We selected 3 kinds of semantic similarities, computed as the cosine between word-type vectors extracted from pre-trained embeddings: *w2v*, which is a 300d Word2Vec Skip-Gram model trained on the English Wikipedia 2018, as found in Wikipedia2Vec (Yamada et al., 2020)[4]; *glv*, which is 300d GloVe model trained on Common Crawl (Pennington et al., 2014)[5]; and *ftx*, which is a 300d FastText model trained on Wikipedia 2017 (Bojanowski et al., 2017).[6] The choice of the similarity between those three will be a hyperparameter to tune. For *SpatialSent*, semantic similarities are computed with the cosine between vectors obtained from a pre-trained sentence embedding model, named *all-mpnet-base-v2* (Reimers & Gurevych, 2019), which is based on BERT (Devlin et al., 2019). All similarities are transformed into dissimilarities with $d_{ij} = \max_{kl} s_{kl} - s_{ij}$. For both methods, we

4 https://wikipedia2vec.github.io/wikipedia2vec/, accessed Jan. 2025.
5 https://nlp.stanford.edu/data/glove.42B.300d.zip, accessed Jan. 2025.
6 https://fasttext.cc/docs/en/english-vectors.html, accessed Jan. 2025.

used the uniform exchange matrix, as it seems to consistently give better results on these tasks.

For each dataset, each task, and each method, the tuning of hyperparameters is done with a grid search on one file of the dataset. This file is selected to be the closest to the typical values found in the dataset in terms of number of tokens, number of sentences, and average topical sequence length. Hyperparameters consist in the choice of $r \in \{5, 10, 15\}$, $\alpha \in \{1, 2, 5, 10, 30\}$, $\beta \in \{5, 10, 50, 100, 200\}$, and $\kappa \in \{0, 0.25, 0.5, 0.75, 1\}$. Moreover, for the *SpatialWord* method, the choice between {*w2v*, *glv*, *ftx*} for the semantic dissimilarity is also tuned.

For the topic clustering task, our methods are compared to Latent Dirichlet Allocation (*LDA*) (Blei et al., 2003) and Non-negative Matrix Factorization (*NMF*) (Arora et al., 2012). However, these methods do not give good results if used as intended, i.e. when using the whole corpus to extract topics. A more efficient way to use them in this task is to split one document into small chunks, and to consider these chunks as different parts containing a mixture of topics. The length of these chunks, which can be selected between fractions $\{\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{5}{20}, \frac{6}{20}, \frac{7}{20}, \frac{8}{20}, \frac{9}{20}, \frac{10}{20}\}$ of document length, is tuned on the same documents as our methods. Using this scheme for these methods allows us to assign different topics to different occurrences of the same word-type, as the probability for a token w to be in group g is $\mathrm{P}(T = g|W = w) \sim \mathrm{P}(W = w|T = g)\mathrm{P}(T = g)$, with a different $\mathrm{P}(T = g)$ for each chunk.

For the text segmentation task, resulting P_k scores for our methods are compared to a random baseline, and scores are reported from *BayesSeg* (Eisenstein & Barzilay, 2008), *GraphSeg* (Glavaš et al., 2016), *TextSeg* (Koshorek et al., 2018), and *Sector* (Arnold et al., 2019) methods. Note that the last two methods, *TextSeg* and *Sector*, are supervised methods and are expected to show better results. Except for *Sector*, a text segmentation and topic classification method, none of these methods give class labels for resulting segments.

3.4 Results

3.4.1 Topic clustering results

Results for the topic clustering task are shown into table 1. The *SpatialSent* consistently performs better than other methods, while *SpatialWord* method is better than *LDA* but less good than *NMF*. Surprisingly, when using some labels, the *SpatialWord* now outperforms *SpatialSent* and gives very strong results. We observe that all method give a low NMI on the MANIFESTO dataset, which is the only annotated dataset permitting a serious validation of the topic clustering task. However, when looking at words defining clusters on a file, even with the *SpatialWord* method as shown in table 2, we can clearly identify pertinent topics using $\mathrm{P}(T = g|W = w)$ or $\mathrm{P}(W = w|T = g)$ (the latter gives the general tone of each topic, which is quite redundant in this case).

	MANIFESTO	WIKI50	CITIES	ELEMENTS	CLINICAL
LDA	12.4	38.0	56.7	48.8	35.0
NMF	19.9	54.7	68.4	61.6	47.6
SpatialWord	14.8	50.2	56.9	45.0	29.0
SpatialSent	**26.9**	**66.6**	**81.5**	**75.9**	**58.1**
SpatialWord, 5%	45.9	72.0	87.2	69.4	67.4
SpatialSent, 5%	30.0	61.9	77.9	76.7	60.4
SpatialWord, 10%	51.7	76.2	91.6	77.0	74.5
SpatialSent, 10%	30.8	66.2	80.2	75.2	68.3

TABLE 1 – Mean NMI results for method × dataset. Higher is better, best results (without considering semi-supervised version of methods) are in boldface.

3.4.2 Text segmentation results

Text segmentation results are found in table 3. Globally, the *TextSeg* seems to perform better than other methods, with the exception of *GraphSeg* for the MANIFESTO dataset (in fact, *TextSeg* was not tested on this dataset) and of *SpatialSent* on the ELEMENTS dataset. If we strictly look at unsupervised methods, we can see that *SpatialSent* generally gives the best results, while *SpatialWord* is in the average. The only other method to also give group labels, *Sector*, gives generally better results than our methods, but these results must be put in perspective because, unlike our methods, it is supervised. When looking at semi-

Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**	**Group 7**
Top 5 word-types regarding P($T = g\|W = w$)						
qadhafi	*america*	*cleaner*	*love*	*withstand*	*bp*	*uninsured*
muammar	*colombia*	*generating*	*tell*	*adversary*	*tsunami*	*counseling*
syrian	*indonesia*	*electricity*	*telling*	*reinforce*	*fissile*	*funds*
prodemocracy	*malaysia*	*cheap*	*splipped*	*involve*	*gm*	*funded*
iraqi	*chile*	*optimizing*	*story*	*reveal*	*nasa*	*refinancing*
Top 5 word-types regarding P($W = w\|T = g$)						
president	*trade*	*energy*	*president*	*president*	*president*	*health*
obama	*president*	*jobs*	*middle*	*continue*	*obama*	*president*
states	*american*	*education*	*back*	*security*	*nuclear*	*care*
support	*economic*	*tax*	*america*	*obama*	*states*	*democrats*
rights	*global*	*businesses*	*work*	*nuclear*	*american*	*support*

TABLE 2 – Word-types defining clusters on the file Democratic 2012 from the MANIFESTO dataset, as found with the *SpatialWord* method (NMI = 11.2).

supervised results, the *SpatialWord* method quickly outperforms all methods, while *SpatialSent* remains about the same, as already seen in the topic clustering results.

	MANIFESTO	WIKI50	CITIES	ELEMENTS	CLINICAL
Random baseline	60.4	61	59	58.5	66.9
GraphSeg	**28.1**	63.6	40	49.1	-
BayesSeg	-	49.2	36.2	35.6	57.8
TextSeg	-	**18.2**	**19.7**	41.6	**30.8**
Sector	-	28.6	21.4	39.2	35.6
SpatialWord	39.6	50.2	49.9	45.4	33
SpatialSent	38.8	43.7	33.9	**28.4**	40.2
SpatialWord, 5%	41.6	33.8	13.2	23.4	27.2
SpatialSent, 5%	38.4	40.2	32.8	30.2	37
SpatialWord, 10%	31.8	24	5.6	10.5	25.5
SpatialSent, 10%	37.3	41	32.1	28.3	30.9

TABLE 3 – Mean P_k results for method × dataset. Lower is better, best results (without considering semi-supervised version of methods) are in boldface.

3.4.3 Method scaling

The method complexity is $\mathcal{O}(n^2 m)$ and does not scale well on large files. However, as seen in section 2.1.5, it is possible to divide a text file into p overlapping blocks of n_b tokens, and proceed sequentially on blocks while transferring predicted labels from the previous block as fixed labels on the next block. We tested this process with *SpatialWord*

on the largest file in our datasets, i.e. the Republican 2020 from the MANIFESTO dataset, which has 25'870 tokens (without stopwords). When using block sizes of n_b, we define each block to have $n_b/2$ overlapping tokens with its predecessor, giving a theoretical complexity of $\mathcal{O}(n_b^2 nm)$. Results are found in figure 2. We see that the computing time is reduced, as well as performances in clustering, as shown by NMI. However, this loss in performances becomes acceptable for largest block sizes. By contrast, P_k seems less affected by the computation on blocks and gives comparable results, even with low block sizes. This block method has not been tested on *SpatialSent*, as the number of sentences is much lower than the number of words and the computing time is reasonable for every document in our datasets.

4 Conclusion

We have presented a very general, classical, formalism which is able to fuzzily cluster textual objects by taking into account a balance between object similarity and position in text. We proposed two methods derived from this formalism: *SpatialWord*, which applies on word-tokens, and *SpatialSent*, operating on sentences. These methods showed good performances for automatically retrieving topics, associated vocabulary, as well as textual segments where these topics appear. Hence, these methods could be used as a new distant reading tool in order to extract topical information on a single document, without any previous training. When compared to state-of-the-art methods on two different tasks, topic clustering and text segmentation, the proposed methods give good results considering they perform both tasks at the same time, in an unsupervised way, and without sharing information across documents. The number of hyperparameters, while permitting these methods to be highly flexible, can however becomes problematic if these methods are applied without knowing ground truth, as no self-validation indices (such as perplexity for topic modeling or inter-group variance for k-means) have been developed for the moment. However, experiments on these datasets on both tasks have already shown some regularities for the studied corpora, and we recommend using *ftx* semantic similarities while setting $r = 15, \alpha = 10, \beta = 100, \kappa = 0.25$ for *SpatialWord*

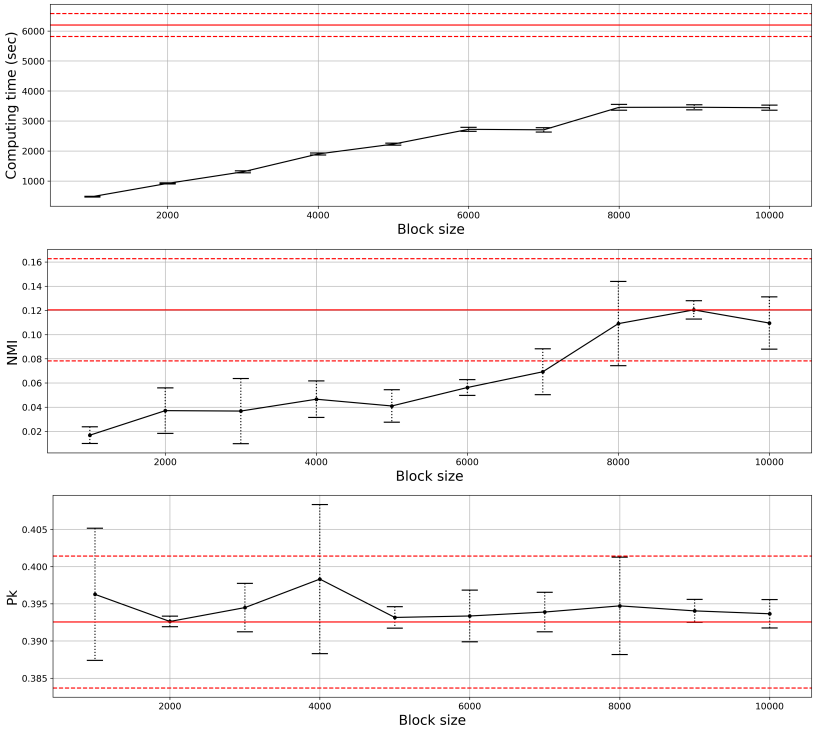


FIGURE 2 – Computing time (left), NMI (middle), and P_k (right) vs the number of tokens in blocks when run on the Republican 2020 file from the MANIFESTO dataset (with 95% CI). Horizontal lines correspond to results when the algorithm is run on the whole file. Results are computed on a single thread, 2.6GHz, i7-9750H CPU.

method, and $r = 5, \alpha = 10, \beta = 100, \kappa = 0.75$ for *SpatialSent* in order to obtain decent results on both tasks at once. Experiments also showed that *SpatialSent* performs consistently better than *SpatialWord* when used without information, which seems obvious as manually annotated segments are generally delimited by punctuation. Surprisingly, *SpatialWord* outperforms greatly *SpatialSent* when having access to some labeled words. While the situation of having direct access to token labels seems unrealistic in real world situations, this could still have applications: a user could provide short lists of typical words defining expected topics (e.g. 10 words for every topic), label corresponding

tokens, and find a pertinent segmentation for his query, as well as the rest of the associated vocabulary. The only remaining difficulty with these methods is a very large computing time. We suggested a way to alleviate this problem, if someone desires to apply them to very large files, but we showed that the performances in the topic segmentation task then decreased. Nevertheless, this should not be problematic if this method is used as an exploratory tool on relatively small corpora, which is the usual setting for a digital humanities researcher.

References

Agarwal, S. & Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Anselin, L. (2010). Local indicators of spatial association-LISA. *Geographical Analysis*, 27(2):93–115.

Arnold, S., Schneider, R., Cudré-Mauroux, P., Gers, F. A., & Löser, A. (2019). Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models – going beyond SVD. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, pages 1–10, USA. IEEE Computer Society.

Bavaud, F., Cocco, C., & Xanthos, A. (2015). Textual navigation and autocorrelation. In Mikros, G. K. & Macutek, J. (Eds.), *Sequences in Language and Text*, pages 35–56. De Gruyter Mouton, Berlin, München, Boston.

Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ceré, R. & Bavaud, F. (2018). Soft image segmentation: on the clustering of irregular, weighted, multivariate marked networks. In *Communications in*

Computer and Information Science, pages 85–109. Springer International Publishing.

Chen, H., Branavan, S., Barzilay, R., & Karger, D. R. (2009). Global models of document structure using latent permutations. In Ostendorf, M., Collins, M., Narayanan, S., Oard, D. W., & Vanderwende, L. (Eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 371–379, Boulder, Colorado. Association for Computational Linguistics.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26–33, USA. Association for Computational Linguistics.

Cressie, N. A. C. (1993). *Statistics for spatial data*. John Wiley & Sons, Inc., New York, NY.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., & Solorio, T. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eisenstein, J. & Barzilay, R. (2008). Bayesian unsupervised topic segmentation. In Lapata, M. & Ng, H. T. (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Glavaš, G., Nanni, F., & Ponzetto, S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 125–130. Association for Computational Linguistics.

Koshorek, O., Cohen, A., Mor, N., Rotman, M., & Berant, J. (2018). Text segmentation as a supervised learning task. In Walker, M., Ji, H., & Stent, A. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., & Weinberger, K. (Eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Nadler, B., Lafon, S., Kevrekidis, I., & Coifman, R. (2005). Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In Weiss, Y., Schölkopf, B., & Platt, J. (Eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., & Daelemans, W. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pevzner, L. & Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., & Wan, X. (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Riedl, M. & Biemann, C. (2012). TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Strehl, A. & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec):583–617.

Tepper, M., Capurro, D., Xia, F., Vanderwende, L., & Yetisgen-Yildiz, M. (2012). Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2001–2008, Istanbul, Turkey. European Language Resources Association (ELRA).

Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., & Weßels, B. (2017). Manifesto project dataset (version 2017b). *Berlin: Wissenschaftszentrum Berlin Für Sozialforschung*.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

A Appendices

A.1 Spatial clustering/semi-supervised classification algorithm

Input: dissimilarity matrix $\mathbf{D}$, exchange matrix $\mathbf{E}$, initial membership matrix $\mathbf{Z}^0$, the set of tagged objects $\mathcal{T}$ (can be empty in case of clustering), hyperparameters α, β, κ, learning parameter $\lambda \in (0, 1]$, stopping threshold ϵ.
Output: The fuzzy membership matrix of objects $\mathbf{Z}$

1 $\mathbf{f} \leftarrow \mathbf{E}\mathbf{e}_n$; // Object weights. $\mathbf{e}_n$ is the size n vector of ones.
2 $\mathbf{Z} \leftarrow \mathbf{Z}^0$; // Initialize membership matrix.
3 $\mathcal{F} \leftarrow 10^9$, $\mathcal{F}^{old} \leftarrow 10^{10}$; // Initialize free energy.
4 **while** $|\mathcal{F} - \mathcal{F}^{old}| > \epsilon$ **do**
5 $\quad \mathbf{Z}^{old} \leftarrow \mathbf{Z}$; // Save old membership matrix.
6 $\quad \mathcal{F}^{old} \leftarrow \mathcal{F}$; // Save old free energy value.
7 $\quad \boldsymbol{\rho} \leftarrow \mathbf{Z}^\top \mathbf{f}$; // Group weights.
8 $\quad \mathbf{F} \leftarrow (\mathbf{f}(\mathbf{e}_n \oslash \boldsymbol{\rho})^\top) \odot \mathbf{Z}$; // Within-group distibutions.[7]
9 $\quad \boldsymbol{\delta} \leftarrow \mathbf{diag}(\mathbf{F}^\top \mathbf{D}\mathbf{F})$; // Group inertias.[8]
10 $\quad \mathbf{c} \leftarrow (\boldsymbol{\rho}^2 - \mathbf{Z}^\top \mathbf{E}\mathbf{Z}) \oslash \boldsymbol{\rho}^\kappa$; // Generalized cut values for groups.[9]
11 $\quad \mathbf{H} \leftarrow \beta(\mathbf{D}\mathbf{F} - \frac{1}{2}\mathbf{e}_n\boldsymbol{\delta}) + \mathbf{e}_n[\alpha\boldsymbol{\rho}^{-\kappa} \odot (\boldsymbol{\rho} - \mathbf{E}\mathbf{Z} \oslash \mathbf{f}\mathbf{e}_m^\top) - \frac{\alpha\kappa}{2}\mathbf{c} \oslash \boldsymbol{\rho}]^\top$; // Matrix $\mathbf{H}$.
12 $\quad \mathbf{Z} \leftarrow (\mathbf{e}_n\boldsymbol{\rho}^\top) \odot \mathbf{Exp}(-\mathbf{H})$; // Unnormalized membership matrix.[10]
13 $\quad \mathbf{Z} \leftarrow \mathbf{Z} \oslash \mathbf{Z}\mathbf{e}_m\mathbf{e}_n^\top$; // Normalize the membership matrix.
14 $\quad \mathbf{Z} = \lambda\mathbf{Z} + (1 - \lambda)\mathbf{Z}^{old}$; // Move it according to the learning rate.
15 $\quad$ **if** $\mathcal{T} \neq \emptyset$ **then**
16 $\quad\quad z_{ig} \leftarrow z_{ig}^0,\ \forall i \in \mathcal{T}, \forall g.$; // Reset tagged objects (if any) to initial values.
17 $\quad$ **end**
18 $\quad \mathcal{F} \leftarrow \beta\mathbf{f}^\top\mathbf{Z}\boldsymbol{\delta} + \frac{\alpha}{2}\mathbf{e}_m^\top\mathbf{c} + \mathbf{e}_n^\top(\mathbf{F} \odot \mathbf{Log}(\mathbf{F} \oslash \mathbf{f}\mathbf{e}_m^\top))\boldsymbol{\rho}$; // Free energy update.
19 **end**
20 `Return(Z)`

7 $\oslash$ and $\odot$ are componentwise (Hadamard) division and multiplication respectively.
8 $\mathbf{diag}()$ gives the vector with the diagonal of the matrix.
9 Powers of vectors are componentwise.
10 $\mathbf{Exp}()$ and $\mathbf{Log}()$ (on line 18) are componentwise.

A.2 Computations of exchange matrices

Uniform exchange matrix : The uniform exchange matrix $\mathbf{E}^{\mathrm{U}(r)}$ can be obtained with the Metropolis-Hastings algorithm from an adjacency matrix $\mathbf{A}^r = (a^r_{ij}) = (\mathbb{1}(|i-j| \leq r]))$, given the stationary distribution $\mathbf{f}$. It reads

$$\mathbf{E}^{\mathrm{U}(r)} = \mathbf{Diag}(\mathbf{f}) - \mathbf{LB} \quad \text{where} \quad \mathbf{B} = (b_{ij}) = \left(\min \left(\frac{f_i a^r_{ij}}{a^r_{i\bullet}}, \frac{f_j a^r_{ji}}{a^r_{j\bullet}} \right) \right)$$

where $\mathbf{Diag}(\mathbf{f})$ is the diagonal matrix containing $\mathbf{f}$ and $(\mathbf{LB})_{ij} := \delta_{ij} b_{i\bullet} - b_{ij}$ the Laplacian of $\mathbf{B}$.

Diffusive exchange matrix : Here, we use a diffusive process from the adjacency matrix $\mathbf{A} = (a_{ij}) = (\mathbb{1}(|i-j| = 1]))$. It gives

$$\mathbf{E}^{\mathrm{D}(t)} = \mathbf{Diag}(\mathbf{f})^{1/2} \mathbf{Exp}(-t\mathbf{\Psi}) \mathbf{Diag}(\mathbf{f})^{1/2}$$

where the $\mathbf{Exp}()$ is the matrix exponentiation and

$$\mathbf{\Psi} := \mathbf{Diag}(\mathbf{f})^{-1/2} \frac{\mathbf{LA}}{\mathrm{tr}(\mathbf{LA})} \mathbf{Diag}(\mathbf{f})^{-1/2}$$

with $(\mathbf{LA})_{ij} = \delta_{ij} a_{i\bullet} - a_{ij}$ the Laplacian of $\mathbf{A}$.

Cahiers du CLSL, n° 69, 2025, pp. 97-109

La pragmatique de corpus comme lieu d'expérimentation des méthodes mixtes : pour une interdisciplinarité focalisée

Jérôme Jacquin
Université de Lausanne
jerome.jacquin@unil.ch

Résumé

La contribution rend compte des développements récents en pragmatique de corpus, champ d'études au sein duquel la pragmatique linguistique tire bénéfice des acquis de la linguistique de corpus et de l'émergence de méthodes mixtes en son sein. Au travers de deux exemples, la contribution plaide pour une interdisciplinarité focalisée où les spécialistes de méthodes informatiques et statistiques permettent indirectement d'enrichir une discipline, la pragmatique, qui trouve ses origines plutôt dans l'analyse détaillée de (collections de) cas.

1 Introduction

En 2013, les sections de linguistique et d'informatique et méthodes mathématiques de la Faculté des lettres de l'Université de Lausanne fusionnent, sous la forme d'une nouvelle entité dénommée « Section des Sciences du Langage et de l'Information ». Cette fusion se réalise par ailleurs à l'occasion d'un redéploiement et d'une redistribution des bureaux de l'ensemble de la Faculté. Ces bouleversements vont entériner une cohabitation à la fois spatiale et institutionnelle allant bien au-delà des échanges scientifiques et personnels ponctuels qui préexistaient à la fusion et qui vont dès lors largement s'intensifier. En 2018, alors que j'occupe désormais un bureau qui jouxte celui de François Bavaud et qui se trouve à une demi-douzaine de mètres de celui d'Aris Xanthos, je dépose un projet auprès du Fonds National Suisse de la recherche (FNS), qui porte clairement la trace de toute une

série de transformations progressives qui se sont opérées en moi et dont une part est précisément imputable à cette cohabitation. Le projet en question, intitulé « Prendre une position épistémique dans l'interaction. Les marqueurs du savoir, du non-savoir et du doute en français » (POSEPI), a été financé par le FNS de 2020 à 2024 [100012_188924]. Il entendait proposer une étude systématique des marqueurs épistémiques (par ex. « peut-être », « il me semble », « c'est pas sûr ») et évidentiels (par ex. « je l'ai vu », « Pierre m'a dit »), tels qu'ils émergent dans un corpus de 28 heures d'interactions naturelles vidéo-enregistrées et documentant des débats politiques et des réunions d'entreprise. Près de 4000 marqueurs ont été annotés (Keck, en préparation, Robin 2024), en suivant un schéma d'annotation exigeant de renseigner des variables catégorielles aussi bien morphosyntaxiques que pragmatiques au sens large, c'est-à-dire énonciatives, interactionnelles et multimodales (direction du regard, gestes cooccurrents ; voir *infra*). Les quelques pages qui suivent abordent trois aspects qui permettent de lier ce projet à la présence concrète et aux intérêts de recherche et de réflexion de François Bavaud. Je commencerai par situer la pragmatique de corpus comme champ d'étude émergeant invitant à adopter des méthodes mixtes au sein d'une discipline, la pragmatique, dont les origines s'ancrent plutôt dans l'analyse détaillée de (collections de) cas, mais qui aujourd'hui bénéficie de plus en plus de notions et outils statistiques. Dans la seconde partie, j'aborderai deux exemples plus concrets à l'occasion desquels les méthodes dites quantitatives et le dialogue direct avec François Bavaud et son équipe ont pu résoudre des défis qui se présentaient : le test d'accord inter-annotateurs sur des variables non discrètes et l'utilisation d'analyses factorielles de correspondances multiples pour articuler de manière outillée les distributions fréquentielles d'ensemble au niveau macro et les analyses séquentielles de cas au niveau micro.

2 La pragmatique de corpus comme champ émergeant

2.1 Développements récents de la linguistique de corpus

La linguistique de corpus a longtemps été associée à l'analyse quantitative de données textuelles, essentiellement écrites (voir par ex. Biber et al. 1998; Svartvik 1992 ; voir également Gries 2010, 2017 sur l'aspect

quantitatif). Depuis une vingtaine d'années, on constate toutefois une double ouverture. La première ouverture concerne un au-delà de la seule quantification et de l'analyse centrée sur des observations statistiques. Cela est rendu possible par un double mouvement. Dans un premier temps, cela se réalise par l'émergence d'une approche plus herméneutique et praxéologique des corpus, qui en font des ressources situées, pilotées par des objectifs de recherche et dont la valeur est fonction des questions que l'on pose et des résultats plus ou moins généraux que l'on cherche à obtenir (voir par ex. Egbert et al., 2020; Mayaffre, 2005; Rastier, 2004). Dans un second temps, on constate un développement de plus en plus soutenu des méthodes mixtes, qui associent des analyses statistiques à des études fines de collections de cas (voir par ex. Angouri, 2010; Riazi, 2016; Teddlie & Tashakkori, 2009). La seconde ouverture de la linguistique de corpus tient aux supports analysés et relève d'un au-delà de la seule scripturalité. Cette ouverture est favorisée par des développements théoriques, analytiques et techniques autour de la multimodalité du langage. On entend par là le fait que les ressources verbales ne constituent qu'une dimension sémiotique pertinente parmi d'autres et qu'une science du langage doit pouvoir intégrer le rôle du corps, qu'il s'agisse de postures, d'expressions faciales telles que la direction du regard ou encore de conduites gestuelles (voir par ex. Bateman et al., 2017; Kress & Van Leeuwen, 2001; Sidnell & Stivers, 2005). Ces développements multimodaux au-delà des corpus uniquement écrits sont aussi imputables à l'émergence, à l'agrégation et à la (encore très relative) mutualisation de corpus vocaux, voire multimodaux (voir par ex. Adolphs & Carter, 2013; Avanzi et al., 2016; Baldauf-Quilliatre et al., 2016). En définitive, les développements récents de la linguistique de corpus s'articulent aujourd'hui très volontiers avec le souhait d'une linguistique basée fondamentalement sur l'usage, en tant que cet usage exige des méthodes et des approches plurielles appliquées à des corpus sémiotiquement complexes. La complexité des corpus ne constitue finalement que le reflet de celle des diverses pratiques plus ou moins ordinaires du langage que ces corpus renseignent.

2.2 Développements récents de la pragmatique

La pragmatique linguistique constitue *grosso modo* l'étude de la relation bidirectionnelle entre les mots et la situation de communication (voir par ex. Bublitz & Norrick, 2011; Huang, 2017; Levinson, 1983). La pragmatique peut donc constituer aussi bien l'analyse de la part du sens des mots qui dépend de facteurs contextuels (par ex. les unités de la deixis comme « je », « ici », « maintenant ») que celle des changements contextuels opérés grâce aux mots (par ex. les actes de langage, les termes à contenus procéduraux, comme les connecteurs, ou encore la gestion de la relation interpersonnelle au travers des ressources relevant de la politesse). Si les origines de la pragmatique se situent principalement en philosophie du langage, puis en linguistique structurale postsaussurienne, on peut noter depuis la deuxième moitié du vingtième siècle une explosion de travaux empiriques. Ces développements empiriques se manifestent aussi bien dans un versant expérimental hérité des travaux en psycholinguistique, que dans un versant plus ethnographique relevant de l'observation et de l'enregistrement de terrains authentiques (sur ces deux tendances, voir par ex. Bublitz et al., 2018). L'étude de la parole en tant que phénomène situé, contextuel, se développe en linguistique puis rencontre, dès les années 1970, l'analyse conversationnelle d'inspiration ethnométhodologique (Sacks et al. 1974; Sidnell & Stivers 2013 ; voir aussi le développement de la « linguistique interactionnelle », p.ex. Couper-Kuhlen & Selting 2018; Selting & Couper-Kuhlen 2001). L'analyse conversationnelle propose non seulement des protocoles éprouvés de récolte et de confection de données interactionnelles (notamment des conventions de transcription), mais aussi une certaine manière d'envisager la posture analytique : il s'agit de s'efforcer d'être au plus près de la dynamique séquentielle des échanges tels qu'ils sont méthodiquement et collaborativement élaborés, pas-à-pas, par les participantes et participants à l'interaction. Ces développements conduisent à la confection de données structurées composées d'enregistrements et de transcriptions documentant une large diversité de contextes linguistiques et socioculturels, aussi bien ordinaires que professionnels et institutionnels (par ex. Arminen, 2005; Berger & Lauzon, 2017; Drew & Heritage, 1992). Ces collections de données documentant des usages

variés sont régulièrement désignées en tant que bases de données, mais aussi corpus, ce qui a posé la question de leur éventuelle ouverture à la linguistique de corpus.

2.3 L'articulation de méthodes dites qualitatives et quantitatives en pragmatique de corpus

De manière analogue à la discussion autour de l'apparent pléonasme de la « socio-linguistique » (par ex. Encrevé, 1977; Labov, 1972), il pouvait sembler aller de soi que linguistique de corpus et pragmatique étaient destinées à se rencontrer du fait de leur affinité avec la notion d'usage et de contexte, notamment autour de l'idée d'indexicalité (par ex. Bar-Hillel, 1954; Duranti & Goodwin, 1992; Levinson, 1983). On aurait même pu imaginer que la pragmatique devienne initiative, voire fédératrice pour la linguistique de corpus ; ou en sens inverse que la linguistique de corpus émerge comme une autre manière d'« intégrer la pragmatique », non pas au niveau théorique du dialogue avec la sémantique, à la suite de Culioli puis d'Anscombre et Ducrot (Anscombre & Ducrot, 1976, p. 8) notamment, mais au niveau plus opérationnel des méthodes et des analyses. Ce n'est de loin pas le cas. Pour reprendre les termes de Rühlemann & Aijmer (2014, p. 1), « for a long time pragmatics and corpus linguistics were regarded as 'parallel but often mutually exclusive' (Romero-Trillo, 2008a, p. 2) ». Dans les années 1980 et 1990, on constate toutefois quelques travaux précurseurs (voir notamment les recensements proposés par Rühlemann & Aijmer (2014); Jucker (2018)), puis une intensification dans les années 2000 (par ex. Romero-Trillo, 2008b) et surtout 2010 (par ex. Romero-Trillo, 2013; Rühlemann, 2019; Weisser, 2018), y compris en francophonie (par ex. Mellet, 2008, sur les connecteurs). C'est en 2014 qu'émerge vraiment l'expression « Corpus Pragmatics » [Pragmatique de corpus] en tant que champ disciplinaire davantage constitué et consolidé, du moins dans la tradition anglo-saxonne (Aijmer & Rühlemann, 2014). À partir de là, non seulement le domaine se développe (voir *supra*), mais la pragmatique prend petit à petit la mesure de cette émergence qui fait bouger les lignes (par ex. Bublitz et al., 2018, part. V). En effet, la pragmatique de corpus invite non seulement à la quantification, mais aussi à l'adoption

de méthodes mixtes (voir aussi *supra*). Cela passe par des analyses statistiques, qui vont de simples fréquences absolues et relatives à des statistiques inférentielles bivariées et multivariées, qu'on exemplifiera *infra* par les analyses factorielles de correspondances multiples[1]. Ces analyses de panoramas d'ensemble, qui permettent de dépasser les apories de l'intuition, de la généralisation abusive et de la saillance des exceptions, sont complétées par des analyses dites « qualitatives », qui visent à proposer des analyses détaillées de cas ou de collections de cas, en considérant divers facteurs contextuels s'accommodant mal d'une annotation massive, tels que les facteurs discursifs, interactionnels ou multimodaux. Cela peut notamment se réaliser par la « next turn proof procedure », c'est-à-dire la prise en compte de la réaction de l'allocutaire dans la fixation du sens de ce qui précède (par exemple, identifier une réponse, permet de situer l'apparition d'une question indépendamment de son formatage syntaxique ou prosodique).

3 Deux exemples d'intégration de méthodes statistiques en pragmatique

J'aimerais développer maintenant deux exemples récents où la collaboration avec François Bavaud et son équipe a pu servir le projet de recherche mentionné en introduction, en développant des méthodes et questions innovantes en pragmatique de corpus de manière à trouver des solutions non pas vraiment sur mesure, mais renseignées et alignées sur l'objectif.

3.1 L'accord inter-annotateurs de variables non discrètes

Le projet de recherche POSEPI a impliqué l'annotation de près de 4000 tokens au travers d'un schéma d'annotation nécessitant de renseigner de très nombreuses variables catégorielles (par ex. des catégories sémantiques, morpho-syntaxiques, discursives, interactionnelles ou multimodales). Il était absolument nécessaire que ces annotations soient suffisamment robustes. Pour ce faire, un guide d'annotation d'une cinquantaine de pages a été rédigé et amélioré au fil de différents tests

1 Comme le précise Jucker, « a quantitative perspective requires a very solid foundation in the preparation of the data base and in the analysis and categorisation of the data » (Jucker, 2018, p. 455). Ce point ne pourra pas être détaillé ici.

d'accords inter-annotateurs sur des échantillons tirés aléatoirement (Jacquin et al., 2022). Ces tests ont été réalisés et interprétés via le Kappa de Cohen (Cohen, 1960; Landis & Koch, 1977). Toutefois, le Kappa de Cohen n'est pas prévu pour des variables non discrètes, comme c'est le cas lorsqu'on doit identifier des portions de matériel continu. Par exemple, dans l'énoncé « il est à la maison je pense », le segment « je pense » constitue un marqueur épistémique portant sur le segment « il est à la maison ». Dans un cas de segmentation comme celui-ci, l'accord inter-annotateur serait probablement assuré. Toutefois, si l'énoncé à considérer est « visiblement lui il est à la maison je pense », cela se complique : « visiblement » fait-il partie ou non de la portée de « je pense » ? Et « lui » ? Le Kappa de Cohen ne fait pas la différence et un écart, quel qu'il soit, entre deux annotateurs mènerait à un désaccord complet ; il n'y a pas de proportionnalité dans l'accord. Le projet MODAL, qui portait sur la modalité à l'oral dans différentes langues (par ex. Nissim & Pietrandrea, 2017; Pietrandrea, 2018), s'est posé la même question. La solution trouvée repose sur la fixation de seuils de chevauchements entre les annotateurs (par ex. 10%, puis 50% et enfin 100%), permettant de diviser le continu en segments discrets et d'ainsi appliquer un Kappa de Cohen classique (Ghia et al., 2016). Une telle solution, bien qu'habile, m'a semblé insatisfaisante du fait de la fixation arbitraire des paliers. Une discussion spontanée avec François Bavaud a permis de poser le problème et de remettre en perspective l'objectif. Elle a conduit à différents échanges où j'ai pu préciser mon idée de mobiliser le *timecode* des annotations respectivement associées aux marqueurs et aux portées, pour calculer des pourcentages d'accord en fonction du degré de recoupement entre les annotations fournies par les deux annotatrices. Ainsi, le résultat du Kappa de Cohen pour chaque token réfléchit le degré d'accord sur les portions annotées respectivement par l'annotatrice A et l'annotatrice B, plutôt que de fixer un seuil qui ferait basculer le recoupement comme un accord ou un désaccord. La formule proposée par François adapte le « coefficient de communauté » de Jaccard (lui-même élaboré dans un premier temps par Grove Karl Gilbert en 1884) et permet ainsi de mesurer le ratio de chevauchement entre deux intervalles temporels en fonction de l'intervalle

total (Bavaud, 2021)[2]. Appliqué à nos données et plus précisément à nos tests d'accord inter-annotateurs, la formule aboutit à un Kappa de Cohen adapté de 0.728 pour les marqueurs et 0.772 pour les portées, ce qui constitue des accords « substantiels » au sens de Landis & Koch (1977). Cette collaboration ponctuelle avec François Bavaud est un exemple où un spécialiste en mathématiques et statistiques aide à traduire en procédure robuste une intuition basée sur un problème pratique émergeant en pragmatique linguistique.

3.2 Méthodes mixtes et identification des (collections de) cas soumis à l'analyse de détails

Le second exemple ne relève pas de l'annotation, mais de l'analyse, et plus précisément de la question relativement délicate de l'articulation entre les distributions statistiques d'ensemble et les collections de cas spécifiques faisant l'objet de l'analyse de détails. En effet, un des risques des méthodes mixtes est que les analyses dites respectivement quantitatives et qualitatives soient disjointes, dans la mesure où le choix des observables soumis aux secondes ne repose pas sur les observations fournies par les premières. Deux méthodes sont ainsi mobilisées dans le cadre du projet POSEPI : la Classification Ascendante Hiérarchique (CAH) et l'Analyse (factorielle) de Correspondances Multiples (ACM). Pour des raisons de place, je vais illustrer l'intérêt de la seconde, mais la plupart des conclusions seront valables pour la première. Les ACM vont consister à sélectionner des variables catégorielles (qualitatives) pertinentes au niveau de la question posée et à étudier et visualiser l'attraction entre elles, sur un plan à deux dimensions au moins (voir par ex. Benzécri & collab., 1973)[3]. Dans l'exemple de la figure 1, l'idée était de voir dans quelle mesure il y a une interaction entre les variables suivantes : le genre d'événement au sein duquel le marqueur a été annoté (s'il s'agit d'un débat public, d'un débat télévisé, ou d'une réunion professionnelle), le type morpho-syntaxique du marqueur (par ex. un

2 Cette stratégie ne doit pas être confondue avec un Kappa de Cohen dit « pondéré » [*weighted*] (Cohen, 1968), qui consiste à travailler le Kappa de manière à nuancer le poids de certaines différences interpersonnelles dans l'annotation.

3 Je remercie ainsi Guillaume Guex, collaborateur de François Bavaud, de nous avoir présenté cette méthode et de nous avoir aidés à l'appréhender sur nos données.

adverbe, ou un introducteur de complétive), la direction entre marqueur et portée (est-ce que le marqueur intervient avant ou après la portée), la position du marqueur au sein de l'unité de construction du tour (au début, au milieu, à la fin) et enfin la dimension d'épistémicité en jeu (modalité épistémique d'une part, évidentialité de l'autre). Sans entrer

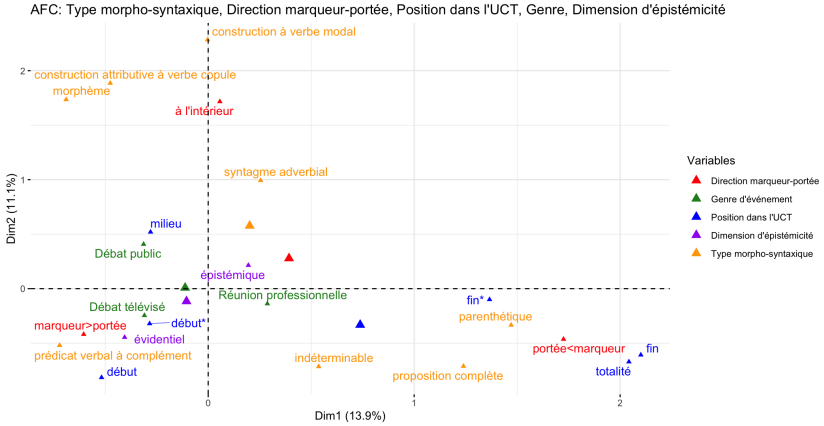


FIGURE 1 – ACM mobilisant cinq variables : le genre d'événement, le type morpho-syntaxique, la direction entre marqueur et portée, la position du marqueur au sein de l'unité de construction du tour et la dimension d'épistémicité en jeu. Graphique réalisé avec R (FactoMineR, factoextra et plotly).

dans le détail, cette ACM permet d'identifier des tendances générales, comme le fait que les positions respectives des marqueurs en fonction des portées (en rouge) sont liées aux types de construction morpho-syntaxiques mobilisées (en jaune) et à la position du marqueur au sein de l'unité de construction du tour de parole (en bleu). Ces tendances donnent lieu à au moins trois groupes (les trois angles du triangle), qui peuvent dans un second temps être analysés en tant que collections de cas à étudier dans le détail de leurs particularités énonciatives, interactionnelles ou multimodales. Cette manière de procéder répond ainsi à une interrogation fréquente au sujet de la manière dont les méthodes dites quantitatives, statistiques, fréquentielles, peuvent s'articuler aux analyses qualitatives, contextuelles, détaillées de (collections de) cas, et donc à la manière dont l'une peut renseigner l'autre.

4 En guise de conclusion : éloge de l'interdisciplinarité focalisée

La présente contribution souhaitait rendre hommage à l'interdisciplinarité focalisée permise et favorisée par la proximité très concrète de spécialistes de domaines différents. Dans le cas présent, pouvoir bénéficier de la présence, de la disponibilité et des compétences de spécialistes en méthodes statistiques et computationnelles a permis et continue de permettre des développements disciplinaires intéressants au sein de la pragmatique, développements qui rejoignent depuis quelques années un intérêt pour la linguistique de corpus et les méthodes mixtes. À plus long terme, on peut imaginer que le perfectionnement continu des logiciels de transcription (semi-)automatisée de paroles spontanées en interaction autorise le défrichage de masses plus importantes de données et donc favorise une plus grande représentativité des données étudiées. L'application d'analyses statistiques non supervisées plus poussées sur des données plus massives devrait aussi permettre de faire émerger des hypothèses plus originales voire imprévues, hypothèses dont les analyses de détails pourraient dans un second temps tester la validité ou simplement la pertinence. Parmi les sous-domaines qui pourraient bénéficier de l'accroissement des données soumises à l'étude et d'une application plus fréquente de méthodes non supervisées, la multimodalité du langage constitue un exemple qui vient rapidement à l'esprit.

Références

Adolphs, S. & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. Routledge, London.

Aijmer, K. & Rühlemann, C. (éd.) (2014). *Corpus pragmatics: A handbook*. Cambridge University Press, Cambridge.

Angouri, J. (2010). Quantitative, qualitative or both? Combining methods in linguistic research. In Litosseliti, L. (éd.), *Research methods in linguistics*, pages 29–49. Continuum, London.

Anscombre, J.-C. & Ducrot, O. (1976). L'argumentation dans la langue. *Langages*, 10(42):5–27.

Arminen, I. (2005). *Institutional interaction: Studies of talk at work*. Aldershot, Ashgate.

Avanzi, M., Béguelin, M.-J., & Diémoz, F. (éd.) (2016). *Corpus de français parlé et français parlé des corpus / Corpus N°15*. Bases, Corpus, Langage - UMR 6039.

Baldauf-Quilliatre, H., de Carvajal, I. C., Etienne, C., Jouin-Chardon, E., Teston-Bonnard, S., & Traverso, V. (2016). Clapi, une base de données multimodale pour la parole en interaction : Apports et dilemmes. *Corpus*, (15).

Bar-Hillel, Y. (1954). Indexical expressions. *Mind*, 63(251):359–379.

Bateman, J., Wildfeuer, J., & Hiippala, T. (2017). *Multimodality, foundations, research and analysis – A problem-oriented Introduction*. De Gruyter Mouton, Berlin, Boston.

Bavaud, F. (2021). Similarité entre intervalles. [Manuscrit non publié]. Université de Lausanne.

Benzécri, J.-P. & collab. (1973). *L'analyse des données. 2 L'analyse des correspondances*. Dunod, Paris.

Berger, E. & Lauzon, V. (2017). *Pratiques interactionnelles en contexte institutionnel / TRANEL (67)*. Université de Neuchâtel, Neuchâtel.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge University Press, Cambridge.

Bublitz, W., Jucker, A. H., & Schneider, K. P. (éd.) (2018). *Methods in pragmatics*. De Gruyter Mouton, Berlin Boston.

Bublitz, W. & Norrick, N. R. (éd.) (2011). *Foundations of pragmatics*, volume 1. Mouton De Gruyter, Berlin.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Couper-Kuhlen, E. & Selting, M. (2018). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press, Cambridge.

Drew, P. & Heritage, J. (éd.) (1992). *Talk at work: Interaction in institutional settings*. Cambridge University Press, Cambridge.

Duranti, A. & Goodwin, C. (éd.) (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge University Press, Cambridge.

Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus*. Cambridge University Press, Cambridge.

Encrevé, P. (1977). Présentation : Linguistique et socio-linguistique. *Langue française*, 34(1):3–16.

Ghia, E., Kloppenburg, L., Nissim, M., Pietrandrea, P., & Cervoni, V. (2016). A construction-centered approach to the annotation of modality. In *Proceedings of the 12th ISO Workshop on Interoperable Semantic Annotation*, volume 29. Harry Bunt, Portoroz.

Gries, S. T. (2010). Methodological skills in corpus linguistics: A polemic and some pointers towards quantitative methods. In Harris, T. & Moreno Jaén, M. (éd.), *Corpus linguistics in language teaching*, pages 121–146. Peter Lang, Frankfurt.

Gries, S. T. (2017). *Quantitative corpus linguistics with R: A practical introduction*. Routledge, Taylor & Francis Group, New York, 2ème édition.

Huang, Y. (éd.) (2017). *The Oxford handbook of pragmatics*. Oxford University Press, Oxford; New York, NY.

Jacquin, J., Keck, A. C., Robin, C., & Roh, S. (2022). Guide d'annotation du projet posepi. Technical Report 1, Université de Lausanne et Fonds National Suisse, Lausanne.

Jucker, A. H. (2018). Introduction to part 5: Corpus pragmatics. In Jucker, A. H., Schneider, K. P., & Bublitz, W. (éd.), *Methods in pragmatics*, pages 455–466. De Gruyter, Berlin.

Keck, A. (en préparation). *Le degré de certitude en français. Étude systématique de la modalité épistémique dans un corpus d'interactions sociales*. Thèse de doctorat, Université de Lausanne. Thèse non publiée.

Kress, G. & Van Leeuwen, T. (2001). *Multimodal discourse*. Arnold, London.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press, Cambridge.

Mayaffre, D. (2005). Rôle et place des corpus en linguistique : Réflexions introductives. Consulté à l'adresse http://www.revue-texto.net/Corpus/Publications/Mayaffre_Corpus.html.

Mellet, S. (éd.) (2008). *Concession et dialogisme : Les connecteurs concessifs à l'épreuve des corpus*. Peter Lang, Berne.

Nissim, M. & Pietrandrea, P. (2017). Modal: A multilingual corpus annotated for modality. In Basili, R., Nissim, M., & Satta, G. (éd.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome*, pages 234–239, Torino. Accademia University Press.

Pietrandrea, P. (2018). Epistemic constructions at work. A corpus study on spoken italian dialogues. *Journal of Pragmatics*, 128:171–191.

Rastier, F. (2004). Enjeux épistémologiques de la linguistique de corpus. Consulté à l'adresse http://www.revue-texto.net/index.php?id=543.

Riazi, A. M. (2016). Innovative mixed-methods research: Moving beyond design technicalities to epistemological and methodological realizations. *Applied Linguistics*, 37(1):33–49.

Robin, C. (2024). *Marquer la source de l'information : Approches interactionnelle, énonciative et multimodale de l'évidentialité en français*. Thèse de doctorat, Université de Lausanne.

Romero-Trillo, J. (2008a). Introduction: Pragmatics and corpus linguistics – a mutualistic entente. In Romero-Trillo, J. (éd.), *Pragmatics and corpus linguistics. A mutualistic entente*, pages 1–10. De Gruyter Mouton, Berlin, Boston.

Romero-Trillo, J. (éd.) (2008b). *Pragmatics and corpus linguistics. A mutualistic entente*. De Gruyter Mouton, Berlin, Boston.

Romero-Trillo, J. (éd.) (2013). *Yearbook of corpus linguistics and pragmatics 2013: New domains and methodologies*. Springer, Dordrecht.

Rühlemann, C. (2019). *Corpus linguistics for pragmatics*. Routledge, New York.

Rühlemann, C. & Aijmer, K. (2014). Corpus pragmatics: Laying the foundations. In Aijmer, K. & Rühlemann, C. (éd.), *Corpus pragmatics: A handbook*, pages 1–26. Cambridge University Press, Cambridge.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Selting, M. & Couper-Kuhlen, E. (éd.) (2001). *Studies in interactional linguistics*. John Benjamins, Amsterdam.

Sidnell, J. & Stivers, T. (éd.) (2005). *Multimodal interaction / Semiotica 156*. Walter de Gruyter, Berlin.

Sidnell, J. & Stivers, T. (éd.) (2013). *The handbook of conversation analysis*. Wiley-Blackwell, Chichester.

Svartvik, J. (éd.) (1992). *Directions in corpus linguistics*. De Gruyter Mouton, Berlin.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE, London.

Weisser, M. (2018). *How to do corpus pragmatics on pragmatically annotated data: Speech acts and beyond*. John Benjamins, Amsterdam.

Cahiers du CLSL, n° 69, 2025, pp. 111-125

Analyse spatiale, émergence en géographie, en archéologie et en histoire

Loïc Jeanson
Université de Lausanne
loic.jeanson@unil.ch

Résumé

Les méthodes dites d'*analyse spatiale* ont été développées après la seconde guerre mondiale et leur adoption transforme profondément les sciences de la terre et du sous-sol (géographie, mais aussi géologie, géodésie, géomorphologie, géotechniques, ...). Cet article présente la genèse, l'évolution d'outils mathématiques et de pratiques en sciences géographiques dans la seconde moitié du XX^{e} siècle. Nous nous attacherons à en suivre les effets par les nouvelles méthodes et nouveaux outils d'analyse en géographie. L'incursion des statistiques rencontrera des résistances fortes qui mèneront à un clivage entre une géographie d'alors, inductive et idiographique, et une géographie nouvelle, nomothétique et déductive. Enfin, en poursuivant jusqu'au XXI^{e} siècle, nous verrons l'essaimage disciplinaire de ces nouvelles méthodes en géographie. L'analyse spatiale, trasnforme aussi l'histoire et l'archéologie par des changements profonds, pratiques, méthodologiques et épistémologiques, parfois réunis aujourd'hui sous le nom de *spatial turn*.

1 Introduction

La polysémie du mot « statistique » peut en compliquer l'emploi, ou plutôt nous force à le définir. Il s'agira qu'elle nous incite à en distinguer les différents sens, afin d'y voir plus clair. D'une part, la statistique et peut-être plus souvent, les statistiques, sont un ensemble de données numériques relatives à des groupes d'individus. D'autre part, la statistique est l'ensemble des méthodes permettant d'analyser ces données. Nous ne parlerons ici que de cette seconde entité : les outils et les méthodes

permettant de traiter les données.

Les divergences de nature entre les différentes méthodes statistiques, leurs développements et transformations, la diversité de leurs contextes d'utilisation et leur ajustement aux spécificités de leurs usages rendent la statistique nébuleuse, compliquant son appréhension en première approche. Et le problème n'est pas neuf.

En 1869, Friedrich Engels livre cette affirmation provocante devant le Congrès International de Statistique : aucun de ses auditeurs, pas même lui, ne serait en mesure de définir ce qu'est la statistique. Il ajoute même avoir récolté 180 définitions différentes. Faute d'avoir pu mettre la main sur cette collection de définitions (même après avoir parcouru les archives d'Engels), Walter Willcox, professeur d'économie à l'Université Cornell et créateur du *US Census Bureau* (Bureau du recensement des États-Unis) fait un constat similaire. Dans un article publié en 1936, il ne recense pas moins de 116 définitions différentes des termes « statistique », « Statistik », ou « Statistics », issue des dernières 50 années de publications scientifiques, soit plus de 2 définitions différentes par an (Willcox, 1936).

Aujourd'hui, on regroupe sous un même chapeau plutôt indifférencié, les données, les méthodes, les techniques. C'est-à-dire qu'on regroupe sous un même titre une grande diversité de pratiques, ajustées à leur objet d'étude et aux types de questions disciplinaires qui mobilisent des méthodes statistiques.

L'usage de la statistique a profondément transformé notre rapport à la mesure, à l'objectivité, notre façon de produire des connaissances dans les sciences expérimentales[1].

La pratique de la statistique transforme les champs disciplinaires, au point, parfois, de les désunir. Parce qu'elle donne des outils nouveaux et finement ajustés aux objets, elle concentre le regard, permet d'étudier plus spécifiquement, de plus près. En cela, elle permet l'ouverture de sous-domaines disciplinaires à part entière. En biologie, par exemple, l'usage de la statistique et du calcul informatique a mené à l'apparition de champs de recherche, appelé les *omics* ou omiques, en français,

1 Sur ce sujet, consulter en particulier les travaux différents de Lorraine Daston (1988, 1989, 1992)

d'après leur suffixe : génomique (genomics), métabolomique (metabolomics), métagénomique (metagenomics), la phénomique (phenomics) et la transcriptomique (transcriptomics). Mais la biologie n'est pas la seule discipline à avoir été profondément bouleversée. Nous verrons dans la suite les transformations qu'a subies la géographie.

2 La géographie, discipline poreuse

La géographie non plus ne se laisse pas cerner facilement. Les géographes, en effet, disputent, de longue date, les frontières qui distinguent leurs pratiques de celles de leurs pairs, tant proches (d'autres géographes) que plus éloignés (les géologues par exemple). Les dissensions ontologiques en géographie sont nombreuses. En 1887, Franz Boas – connu aujourd'hui comme un des fondateurs de l'anthropologie américaine – investigue une étrangeté :

> It is a remarkable fact, that, in the recent literature of geography, researches on the method and limits of that science occupy a prominent place. Almost every distinguished geographer has felt the necessity of expressing his views on its aim and scope, and of defending it from being disintegrated and swallowed up by geology, botany, history, and other sciences treating on subjects similar to or identical with those of geography (Boas, 1887, p. 901).

Aux XVIII^{e} et XIX^{e} siècles, ainsi que dans la première moitié du XX^{e} siècle, produire une géographie, c'est écrire, décrire les espaces explorés ou étudiés. Ainsi, les géographes proches de la botanique sont à même de décrire les climats, la diversité des types d'éco-systèmes. D'autres, les géographes proches des géologues, décrivent la nature des sols, des roches, des paysages par la géomorphologie, etc. Cette grande porosité disciplinaire permet de réunir sous la notion de *géographie* des pratiques et des questions très différentes, composant la discipline d'un grand nombre de facettes. En retour, cette porosité disciplinaire rend la géographie mal définie. On retrouve ce constat d'indéfinition de façon répétée et systématique au travers du XX^{e} siècle (Dryer, 1905; Kropotkin, 1885; de La Blache, 1913; Sorre, 1948, etc.).

Par cette approche descriptive, la géographie s'intéresse aux spécificités, aux *unicum*. Elle dit l'existant et les caractéristiques des localismes. En cela, elle est idiographique, accumulant des cas uniques. Mais l'emploi de méthodes statistiques par certains géographes, de concert avec leurs homologues géologues, naturalistes ou sociologues, annonce des changements profonds pour la discipline.

2.1 La révolution quantitative

En 1948, l'université d'Harvard ferme son département de géographie (Smith, 1987). Cette décision, ébranle la communauté des géographes aux États-Unis et exprime le malaise ancien et les critiques d'indéfinition disciplinaire récurrentes que subit la géographie. La géographie se trouve à un tournant, et la direction d'Harvard doute qu'elle réussira à le prendre, celui de l'emploi des méthodes statistiques, soutenues par l'usage des premiers ordinateurs.

Disons cependant tout de suite que le recours aux ordinateurs n'implique pas forcément l'emploi des méthodes statistiques. L'informatique à des des fins académiques apparaît couramment aux États-Unis dans les années 1940 et en Europe dans les années 60. En 1967, alors que l'informatique n'est désormais plus nouvelle dans les cercles académiques, Torsten Hägerstrand, professeur de géographie à Lund distingue trois fonctions de l'informatique pour la géographie (Hägerstrand, 1967) :

— une fonction descriptive, permettant de produire des cartographies descriptives, en positionnant sur un fond de carte des éléments directement grâce à leur géoréférencement.
— une fonction analytique, permettant d'évaluer des relations spatiales, des relations entre lieux, en calculant des indices, des corrélations entre phénomènes, en calculant des frontières à partir de certaines règles, en classifiant des lieux ou des régions, en appliquant la théorie des cribles, ...
— une fonction de simulation, afin d'employer des modèles pour tenter de reproduire des phénomènes observés ou de tester les hypothèses d'évènements de nature géographique.

Les géographes eux-même affirment des points de vue très différents sur le recours aux méthodes mathématiques, que traduiront des diver-

gences pratiques durables. Si certains, comme John Cole, cherchent à montrer comment la formalisation mathématisation de concepts géographiques permet des approches et des résultats nouveaux (Cole, 1969), d'autres témoignages plus sceptiques (p. ex. Stamp, 1957 ou Gould, 1970) révèlent des divergences établissant deux pratiques de la géographie : d'une part, une géographie idiographique, descriptive, attachée aux évènements pour leur unicité et d'autre part une géographie nomothétique, analytique, quantitative, attachée à mettre au jour des règles plus générales à partir de collection d'évènements.

Deux citations synthétisent ces divergences. D'un côté Sauer, professeur de géographie à Berkeley (il devient émérite en 1957) affirme en 1956 :

> A geographer is any competent amateur [...] we may leave most enumeration to census takers [...] to my mind we are concerned with processes that are largely non-recurrent and involve time spans beyond the short runs available to enumeration (Sauer, 1956, pp. 291ff.).

Face à lui, les mots attribués au philosophe et logicien Whitehead, largement repris par des géographes déployant des approches statistiques : « To see what is general in what is particular and what is permanent in what is transitory » (repris dans Gould, 1979, p. 140).

Malgré la co-existence de regards différents, les méthodes quantitatives gagnent globalement du terrain. Si certains chercheurs décident d'éviter de les employer, entre la fin des années 1950 et la fin des années 1960, tous les départements de géographie intègrent des enseignements relatifs à ces méthodes quantitatives et statistiques (Lavalle et al., 1967).

2.2 Quantitatif et spatial

Malgré des résistances, les méthodes quantitatives gagnent la géographie à partir des années 1950. Johnston et al. décrivent (Johnston et al., 2019) :

— des changements philosophiques (comment la géographie quantitative adopte les méthodes scientifiques d'autres sciences ayant recours à la statistique) ;

— l'apparition du concept d'ordre spatial (il ne s'agit plus de regarder des relations verticales, des populations avec leur environnement, mais des relations horizontales : *spatial interaction, contacts and movements between places* ;
— le besoin d'intégrer pleinement les aspects quantitatifs à la géographie. C'est-à-dire que l'apprentissage de la statistique ne soit pas seulement un moyen d'assurer leur emploi correct, mais bien que leur adoption soit un élément d'une transformation plus large des pratiques géographiques.

Les géographes travaillant aux frontières de différentes disciplines (démographie, écologie, sciences de l'environnement, de la Terre, ...), emploient différentes méthodes statistiques. Mais s'il ne faut garder qu'un aspect propre à la géographie, c'est bien l'étude des espaces, c'est pourquoi, dans la suite de notre propos, nous nous concentrons seulement sur l'autocorrélation spatiale.

Avant 1968, elle a aussi été appelée « dépendance spatiale », ou « interdépendance spatiale », mais aussi « association spatiale », « interaction spatiale », ...

Griffith (1992) présente l'autocorrélation spatiale comme pouvant revêtir les neuf définitions suivantes :

— la corrélation d'une variable avec elle-même provenant de l'ordonnancement géographique des données ;
— un instrument de description de la nature et de l'intensité d'une structure spatiale ;
— un indicateur de la quantité d'information latente contenue dans les données spatialisées, en particulier l'information qui s'avère toujours négligée dans les estimations statistiques classiques quand elles sont appliquées aux séries de données spatiales ;
— un outil permettant de repérer l'existence de variables significatives, mais non prises en compte dans le modèle ;
— un substitut à des données manquantes ;
— un obstacle pour l'application des méthodologies statistiques conventionnelles à des séries de données spatiales ;
— un indicateur du bien-fondé d'une partition spatiale, voire un artefact introduit par les frontières ;

— un mécanisme d'un processus spatial ;
— enfin, un effet de redistribution sur les lieux avoisinants.

Les effets de l'autocorrélation spatiale sont identifiés dès 1914 (Student, 1914) mais il faut attendre la moitié du XX^e^ siècle pour que Moran (1948, 1950) et Geary (1954) définissent deux indices pour mesurer l'autocorrélation spatiale. Ainsi, il devenait possible de mesurer si la position de groupes de valeurs similaires ou dissimilaires, produisent des motifs qui, sans référentiel géographique, seraient inexplicables.

Michael Dacey semble être le premier à proposer une mesure véritablement adaptée à la géographie (Dacey, 1966, 1968), c'est-à-dire qui dépasse les limites de l'invariance topologique propres aux mesures de Moran et Geary. Dans sa proposition, il pondère ainsi la contribution de chaque paire (les aires des zones i et j) par leur proportion de la zone étudiée ($a(i)$) multipliée par proportion de la frontière de i qui est commune entre i et j, ($b(i;j)$).

A sa suite, de nombreux géographes et mathématiciens formuleront différentes façons d'intégrer des pondérations propres à décrire et modéliser correctement les effets géographiques (p. ex. Bavaud, 2013; Getis, 2008; Griffith, 1996).

La pluralité des méthodes adaptées aux enjeux spatiaux développées par les géographes et mathématiciens est communément regroupée sous le terme d'analyse spatiale. En plusieurs décennies de développement des méthodes et de contexte de leur application, les géographes ont un recul sur leurs apports (Oliveau, 2010) et montrent une grande dextérité dans l'ajustement et la mathématisation de ces méthodes (Le Gallo, 2002).

3 L'archéologie et l'histoire s'emparent de ces méthodes

L'histoire et l'archéologie aussi se situent en interface disciplinaire avec la géographie. Avec celle-ci, elles partagent des terrains, des sources, des outils (cartographiques *a minima*). Les développements réalisés en géographie n'ont pas immédiatement été portés dans ces disciplines voisines, mais avec quelques dizaines d'années d'écart.

3.1 L'analyse spatiale gagne l'archéologie

L'analyse spatiale se base sur les trois fonctions de la statistique informatisée explicitées précédemment (visualiser, analyser, simuler). Selon les sites et les problématiques de recherche, les archéologues cherchent des outils pour un usage proche de celui des géographes : visualiser, analyser et tester des hypothèses par rapport à des espaces documentés, à partir d'informations (souvent très) lacunaires. En cela l'analyse spatiale (autrement dit le recours à la statistique et à l'informatique sur des données spatialisées) trouve en archéologie un terrain propice à son utilisation. Les méthodes issue de la statistique (analyse multivariée, autocorrélation, clustering, classification, calcul de distances, de cartes thermiques, ...) appliquées à l'archéologie permettent de mettre en lumière des relations spatiales au travers des échelles de regard : intra-site, inter-site, à l'échelle de territoires, d'aires culturelles, en suivant des circulations, etc. On trouve de nombreux exemples. Prenons-en deux, séparés de plusieurs années, afin de mesurer combien l'adoption des méthodes statistiques se généralise, mais aussi combien le développement logiciel (les systèmes d'information géographiques – SIG), ainsi que le développement de langages de programmation permettant la modélisation et le calcul (en particulier R et Python dans les usages courants d'archéologues aujourd'hui), facilitent ces usages.

Pour suivre le changement de rapport à ces approches, commençons par un article intermédiaire (1984), de Jan F. Simek, professeur d'archéologie à l'Université du Tennessee, spécialiste des habitats paléolithiques et néolithiques dans des grottes et abris sous roche. La figure 1 réunit en synthèse des méthodes statistiques utilisées pour analyser la couche V du site nommé *Le Flageolet I* en Dordogne, en France.

Les analyses de Simek utilisent des méthodes statistiques simples, afin de classifier et réduire le volume de données issues des fouilles. La mathématique utilisée n'est pas très poussée, mais les clusters et les zones sont spatialisés dans un plan 2D de référence. Nous sommes alors en 1984, et si l'accès à l'informatique n'est pas plus limité pour les géographes que pour les archéologues, la connaissance de la pertinence des méthodes et de leurs usages doivent traverser les poreuses parois disciplinaires.

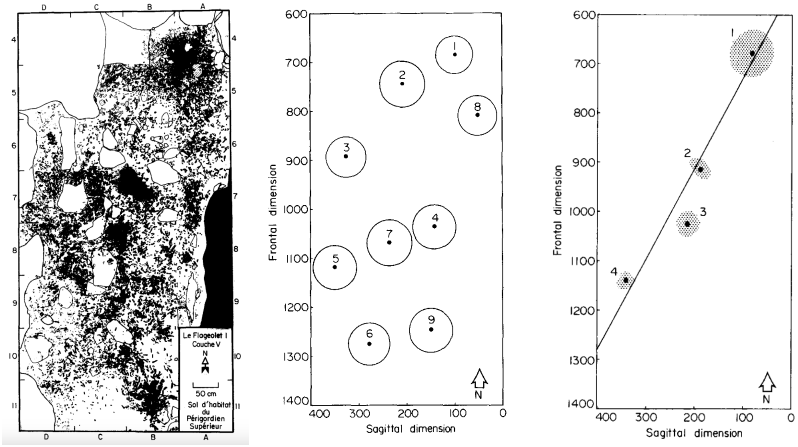


FIGURE 1 – A gauche, localisation des trouvailles archéologiques de la couche V du site de fouille *Le Flageolet I* ; au centre, les résultats de la classification par k-means à partir de typologie des outils identifiés et caractérisés parmi les objets trouvés sur cette même couche V ; à droite les zones 1, 2 3 et 4 montrent un centre et une étendue de localisation des éléments archéologiques (hors objets), la ligne indique la régression par le centre des zones. Ces illustrations, reproduites avec autorisation, viennent de Simek (1984).

Comme second exemple, regardons l'usage fait à une tout autre échelle, bien plus grande. Nous observons alors des sites de fouilles répartis autour des frontières entre Belize, le Guatemala, le Honduras et le Mexique. Luke S. Premo, anthropologue s'intéresse aux dynamiques d'occupation du territoire de populations Maya, au travers de l'analyse de 47 sites. Dans cet article, il montre comment le recours aux indices de Morand et de Geary lui permet de distinguer des dynamiques locales tout en continuant à regarder l'ensemble des sites à une échelle globale. La figure 2 montre la position des sites et les différents niveaux de gris donnent les datations extrêmes qui ont pu être attribuées aux sites. On peut les voir représentés sur la figure 3. Le recours à l'autocorrélation locale, par les indices de Morand et de Geary rend possible la visualisation de dynamiques locales de grain plus fin, lui permettant d'approfondir sa compréhension de l'occupation du territoire et des dynamiques humaines passées.

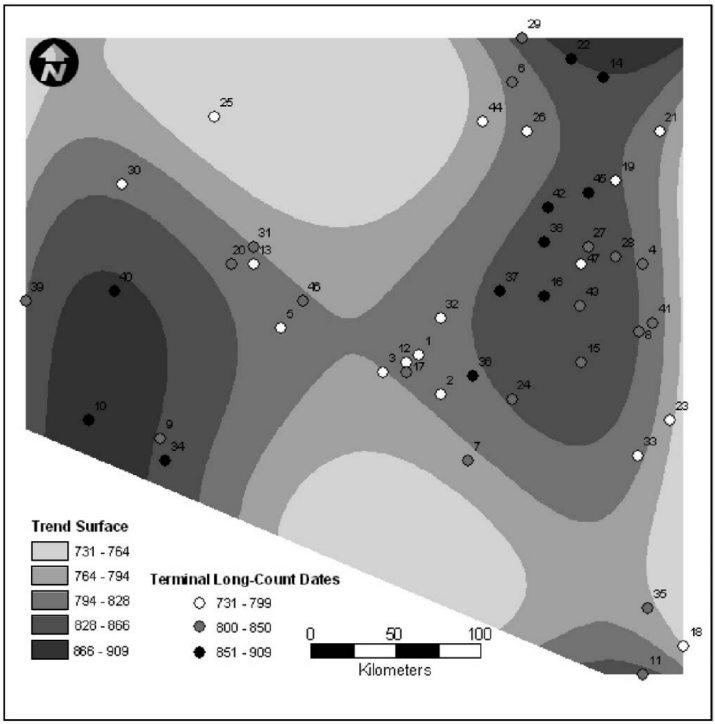


FIGURE 2 – Cette carte (très simplifiée) montre la position et les datations extrêmes des 47 sites Maya étudiés par Premo. Cette illustration, reproduite avec autorisation, vient de Premo (2004).

Il ne s'agit pas ici de produire le récit linéaire, de tracer une direction à suivre ou une perspective unique de progrès. Il s'agit plutôt de montrer la diffusion des méthodes, les transformations épistémologiques, accélérées par un contexte socio-technique d'accès sans cesse plus facile aux langages, aux logiciels et moyens matériels du calcul.

3.2 Le *spatial turn* s'empare de l'histoire, et des histoires spécialisées

Les historien·nes approchent également l'espace au travers de leurs questions propres, mais sous l'influence de géographes, les enjeux

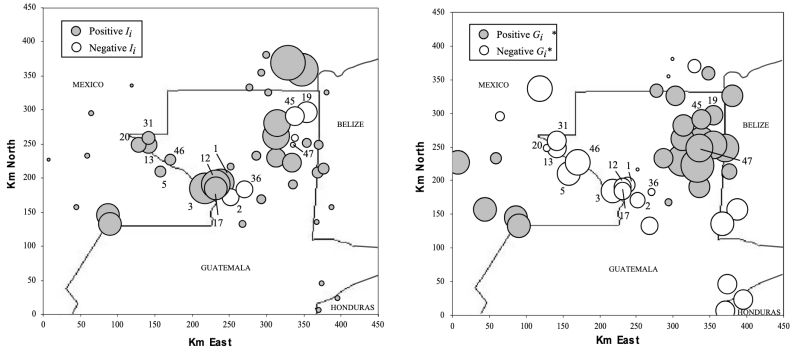


FIGURE 3 – Cette carte (très simplifiée) montre la position et les datations extrêmes des 47 sites Maya étudiés par Premo. Cette illustration, reproduite avec autorisation, vient de Premo (2004).

spatiaux gagnent du terrain dans leurs activités depuis les années 1990 environ (on peut noter en particulier Edward Soja, très influent dans le monde anglophone, notamment avec Soja, 1989).

Les historien·ne·s des sciences et des techniques étaient peut-être déjà confronté·e·s aux enjeux spatiaux de l'étude de leurs objets, dans l'approche de leurs acteurs. Ce furent en tous cas les premièr·e·s à incarner ce *spatial turn* en histoire. Ils·elles cherchaient à « spatialiser l'histoire », « localiser la culture », « situer la rationalité » (pour reprendre les catégories de Soja). Très associé à la révolution numérique, ce *tournant spatial* s'empare des SIG (Juvan, 2015), mais avant que leur usage ne soit véritablement démocratisé par l'existence de logiciels open source bien aboutis. Par exemple, dès 1991, la revue *Science in Context*, créée en 1987, publie un numéro dédié à l'étude du « lieu de savoir » (*the place of knowledge*. L'introduction de ce numéro spécial, par Adi Ophir et Steven Shapin donne une bonne idée du regard des historiens des sciences sur le rapport aux espaces (Ophir & Shapin, 1991).

Entre les années 1990 et aujourd'hui, les travaux d'historien·ne·s intégrant l'analyse spatiale se multiplient. Ils accompagnent les tendances disciplinaires à l'œuvre. La micro-histoire trouve un pendant dans la

micro-spatial perspective (De Vito, 2019). L'étude de la circulation des idées, des paradigmes ou des concepts y trouve également un fort écho (p. ex. dans le programme de recherche CIRMATH, Peiffer et al., 2020). Il en va de même pour les études *aréales* à la frontière entre histoire et géographie. Différents consortium réunissant historien·ne·s et archéologues, voient le jour, permettant le partage de la connaissance des méthodes et la formation à leur utilisation (p. ex. le consortium Athéna, Thibault, 2018). Enfin, depuis une décennie environ, l'histoire de l'urbanisme et de l'architecture a recours largement aux méthodes statistiques d'analyse spatiale. Plusieurs membres du consortium *Paris Time Machine*, par exemple, ont produit des résultats numériques utiles pour les historiens. Citons, en particulier les travaux réalisés entre l'IGN et l'EHESS, ayant permis de géocoder les adresses sur des cartes anciennes de Paris, puis de cartographier les activités et commerces actifs, en recoupant ces adresses avec celles indiquées dans les annuaires du commerce de plusieurs décennies, dont les informations ont été extraites automatiquement (Abadie et al., 2023). On pourra aussi citer l'approche complémentaire, à échelle réduite mais avec un grand détail, réalisée à l'Institut National d'Histoire de l'Art sur l'étude de quelques rues du quartier Richelieu à Paris (Duvette et al., 2023).

4 Conclusion

Si la révolution spatiale et informatique a commencé en géographie dans les années 1960, elle continue à se déployer, tant du point de vue des méthodes que de leurs applications, en géographie, en archéologie, en histoire, etc.

Par ce rapide tour d'horizon, nous espérons avoir pu montrer la pertinence qu'ont trouvé les méthodes statistiques et comment les travaux mathématiques développant des méthodes nouvelles, ou les ajustant à des questions et des approches spécifiques, transforment les pratiques scientifiques, les types de résultats produits, les stratégies d'analyse et les moyens de la preuve mobilisés par les chercheuses et chercheurs. La démocratisation des moyens matériels et logiciels du calcul contribue encore à leur adoption.

La pertinence de ces méthodes et de leurs usages nous révèle leur

utilité dans les enseignements en géographie, en histoire, en archéologie, ou de façon générale dans les sciences humaines et sociales. Si les approches qualitatives et idiographiques restent inévitables pour approcher certains questionnements, les approches quantitatives nous donnent des moyens d'analyse, de modélisation et de visualisation indispensables à d'autres façons de faire la géographie, l'histoire ou l'archéologie. En cela, l'hybridité des approches livre des résultats étayés par des moyens multiples, permettant de croiser et de cumuler les preuves. Le *spatial turn* et l'analyse statistique spatiale ont encore de beaux jours devant eux !

Références

Abadie, N., Baciocchi, S., Bernard, C., Carlinet, E., Chapron, P., Chazalon, J., Chen, Y., Cristofoli, P., Duménieu, B., Fernandez, M., et al. (2023). Soduco : croisement de sources géo-historiques pour l'étude de l'évolution de Paris de 1789 à 1950. In *Conférence interdisciplinaire : Extraction, traitement et visualisation de données complexes en géographie (XVIIIe siècle-XIXe siècle)*.

Bavaud, F. (2013). Testing spatial autocorrelation in weighted networks: the modes permutation test. *Journal of Geographical Systems*, 15(3):233–247.

Boas, F. (1887). The study of geography. *Science*, 9(201):137–141.

Cole, J. P. (1969). Mathematics and geography. *Geography*, 54(2):152–164.

Dacey, M. F. (1966). A county-seat model for the areal pattern of an urban system. *Geographical Review*, 56(4):527–542.

Dacey, M. F. (1968). An empirical study of the areal distribution of houses in Puerto Rico. *Transactions of the Institute of British Geographers*, 45:51–69.

Daston, L. (1992). Objectivity and the escape from perspective. *Social Studies of Science*, 22(4):597–618.

Daston, L. J. (1988). *Fitting numbers to the world: The case of probability theory*. University of Minnesota Press, Minneapolis.

Daston, L. J. (1989). Compte-rendu du livre « The history of statistics: The measurement of uncertainty before 1900 » de Stephen M. Stigler. *The Journal of Modern History*, 61(1):135–137.

de La Blache, P. V. (1913). Des caractères distinctifs de la géographie. *Annales de Géographie*, 22(124):289–299.

De Vito, C. G. (2019). History without scale: The micro-spatial perspective. *Past & Present*, 242(Supplement_14):348–372.

Dryer, C. R. (1905). What is geography? *Journal of Geography*, 4(8):348–360.

Duvette, C., Jeanson, L., Kervegan, P., Prudhomme, C., Gain, J., Dasilva, E., & Baranger, L. (2023). « La marque du lieu » dans le « quartier Richelieu ». In *Humanistica 2023*.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146.

Getis, A. (2008). A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis*, 40(3):297–309.

Gould, P. (1970). Is statistix inferens the geographical name for a wild goose? *Economic Geography*, 46:439–448.

Gould, P. (1979). Geography 1957–1977: the Augean period. *Annals of the Association of American Geographers*, 69(1):139–151.

Griffith, D. A. (1992). What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics. *L'Espace géographique*, 21(3):265–280.

Griffith, D. A. (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer/Le Géographe canadien*, 40(4):351–367.

Hägerstrand, T. (1967). The computer and the geographer. *Transactions of the Institute of British Geographers*, 42:1–19.

Johnston, R., Harris, R., Jones, K., Manley, D., Wang, W. W., & Wolf, L. (2019). Quantitative methods I: The world we have lost–or where we started from. *Progress in Human Geography*, 43(6):1133–1142.

Juvan, M. (2015). From spatial turn to GIS-mapping of literary cultures. *European Review*, 23(1):81–96.

Kropotkin, P. A. (1885). What geography ought to be. *The Nineteenth Century*, 18:940–956.

Lavalle, P., McConell, H., & Brown, R. G. (1967). Certain aspects of the expansion of quantitative methodology in American geography. *Annals of the Association of American Geographers*, 57(2):423–436.

Le Gallo, J. (2002). Econométrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaire. *Economie & prévision*, 155(4):139–157.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Oliveau, S. (2010). Autocorrélation spatiale : leçons du changement d'échelle. *L'Espace géographique*, 39(1):51–64.

Ophir, A. & Shapin, S. (1991). The place of knowledge a methodological survey. *Science in Context*, 4(1):3–22.

Peiffer, J., Gispert, H., & Nabonnand, P. (2020). De l'histoire des journaux mathématiques à l'histoire de la circulation mathématique. *Cahiers François Viète*, III-9:123–153.

Premo, L. (2004). Local spatial autocorrelation statistics quantify multiscale patterns in distributional data: an example from the Maya Lowlands. *Journal of Archaeological Science*, 31(7):855–866.

Sauer, C. O. (1956). The education of a geographer. *Annals of the Association of American Geographers*, 46(3):287–299.

Simek, J. F. (1984). Integrating pattern and context in spatial archaeology. *Journal of Archaeological Science*, 11(5):405–420.

Smith, N. (1987). "Academic war over the field of geography": The elimination of geography at Harvard, 1947–1951. *Annals of the Association of American Geographers*, 77(2):155–172.

Soja, E. W. (1989). *Postmodern geographies: The reassertion of space in critical social theory*. Radical Thinkers. Verso, London New York.

Sorre, M. (1948). Fondements de la géographie humaine. *Cahiers internationaux de sociologie*, 5:21–37.

Stamp, L. D. (1957). Geographical agenda: A review of some tasks awaiting geographical attention: Presidential address. *Transactions and Papers (Institute of British Geographers)*, 23:1–17.

Student (1914). IV. The elimination of spurious correlation due to position in time or space. *Biometrika*, 10(1):179–180.

Thibault, F. (2018). Alliance athéna, activity report 2017 [rapport de recherche]. Alliance Athéna.

Willcox, W. F. (1936). Definitions of statistics. *Revue de l'Institut International de Statistique/Review of the International Statistical Institute*, 3(4):388–399.

Cahiers du CLSL, n° 69, 2025, pp. 127-142

On machine learning from environmental data

Mikhail Kanevski
University of Lausanne
mikhail.kanevski@unil.ch

Abstract

The application of machine learning (ML) algorithms for analyzing, modelling, and visualizing geospatial data has seen remarkable growth in recent years. Many ML techniques have proven to be both effective and efficient in addressing complex challenges within the geosciences. Selected topics on adaptation and application of ML to environmental data are briefly discussed in this research.

1 Introduction

This paper focuses on key aspects of applying machine learning to the analysis and modelling of spatial environmental data. It emphasizes a comprehensive methodology that encompasses from data collection (monitoring network analysis, design and redesign), intelligent exploratory data analysis and visualization, via ML models training and evaluation to understanding and communication of the results for informed decision making (Kanevski & Maignan, 2004; Kanevski et al., 2009). The main components of this methodology are illustrated in figure 1 (p. 132), with further details provided in the subsequent sections.

In practice, ML should be widely utilized across all phases of data-driven modelling. It facilitates comprehensive data exploration, selection of relevant variables, visualization of high-dimensional and large datasets, pattern recognition, adaptive modelling, and result interpretability. Numerous excellent books cover the concepts and theories of ML algorithms, as well as their applications across various domains (see, for example Bishop & Bishop, 2024; Cherkassky & Mulier, 2007;

Hastie et al., 2021; Haykin, 2009). Additionally, an abundance of resources on ML programming and modern open-access packages is now available, further accelerating development and broadening the user community (James et al., 2021; Kuhn & Johnson, 2013).

Recently, a new and rapidly evolving field of research, *geospatial data science*, has emerged as an interdisciplinary domain that integrates geoinformatics, geostatistics, machine learning, network science, and more, see, for example (Gaur et al., 2023; Pebesma & Bivand, 2023). This field encompasses a broad range of applications across diverse domains, including geography and urban planning, earth sciences, environmental science, meteorology and climate science, ecology, and beyond.

Geospatial data present several significant challenges due to their inherent characteristics. Unlike traditional datasets, they often exhibit spatial autocorrelation (Jemeljanova et al., 2024; Kattenborn et al., 2022), meaning that observations are not independently and identically distributed (i.i.d.), which complicates standard statistical and machine learning approaches (Linnenbrink et al., 2023). Additionally, spatial clustering and preferential sampling can introduce biases, affecting model development, evaluation, and the definition of validity domains (Brus, 2022; Meyer & Pebesma, 2021; Schratz et al., 2019).

Moreover, causal analysis in spatial data is particularly challenging due to confounding spatial dependencies and intrinsic uncertainties, making robust inference difficult (Akbari et al., 2023; Gao et al., 2022).

The interpretability of ML-based data analysis and predictions is also becoming increasingly critical in geoscience applications, particularly in studies related to climate, environmental risks, natural hazards, and renewable energy assessments (Jiang et al., 2024).

Addressing these challenges requires the development of specialized methodologies to ensure accurate modelling, reliable predictions, and informed decision-making in geospatial applications.

Prof. F. Bavaud has made important contributions to geospatial data science, particularly in developing innovative algorithms and methodologies (Bavaud, 2009, 2014, 2024; Guex et al., 2023). His work effectively combines theoretical rigour with practical applications, bridging

the gap between fundamental research and real-world challenges.

1.1 Why machine learning?

The application of machine learning in environmental data analysis offers numerous benefits that enhance data processing and modelling while addressing complex environmental problems. Here are several key advantages:

- ML algorithms are powerful *universal non-linear* modelling tools. They can model data with high precision, adapting to a wide range of structures and variability.
- ML exhibit strong generalization capabilities: they make accurate predictions on new unseen (testing) data.
- ML models can handle big and complex data sets from diverse sources.
- ML techniques perform well with high-dimensional data: many algorithms have been specifically developed for non-linear dimensionality reduction and feature selection.
- ML algorithms can be integrated into automatic spatio-temporal environmental data processing and monitoring.
- ML models have already demonstrated their efficiency and usefulness in numerous environmental applications (weather, climate, natural hazards, pollution, renewable resources, ecology, biodiversity, etc.).

Despite their significant success, the use of ML algorithms can encounter various challenges. ML heavily depends on the quality and quantity of data; in real-world applications, training, selecting, and evaluating ML models are non-trivial tasks. Additionally, making predictions and forecasts that account for uncertainties is often difficult. The interpretability and explainability of ML models remain active areas of contemporary research (Jiang et al., 2024). It is also crucial to note that the effective application of ML requires a deep understanding

of algorithms, a grasp of underlying assumptions, and collaboration with domain experts.

1.2 Model-centric and data-centric machine learning

In the field of artificial intelligence (AI), there are two primary approaches to improve the results: model-centric and data-centric, often called MCAI and DCAI. Each focuses on different aspects of ML modelling and has its particular advantages.

Model-centric focuses on model architecture and strengthening the algorithms, in particular on model selection and evaluation, hyperparameters tuning, optimization techniques, combining models, ensemble learning (Bartz et al., 2023; Bishop & Bishop, 2024; Hastie et al., 2021; Montavon et al., 2012). Selection of ML model appropriate to data and objectives can help to achieve state-of-the-art results in many applications. Until recently, model-centric modelling has dominated in developments.

Nowadays, the role of data in ML modelling is promoted by the fast developing concept of data-centric ML (Mahalle et al., 2024). This demonstrates an important shift from model strengthening to data quality and reliability.

In data-centric approach the focus is on *systematic* improvement of data quality and quantity in the process of ML training and evaluation. Data are considered as a dynamic object, while model is usually fixed. Fundamentally there are two possibilities: either improving training dataset using available data or collection/simulation (data augmentation) of additional data.

There are many tools and techniques that are a part of data-centric approach: missing values treatment, outliers/anomalies detection and removal, data validation and correction of errors, feature engineering and data reduction (features and/or instances selection), active learning (guided data selection), interactive visualization. All these techniques improve the quality and quantity of training and testing data sets giving rise to better predictions and reducing errors.

One of the first question in the analysis concerns data representativity, i.e., how available data represent the phenomena under study? In

spatial statistics this is mostly related to data clustering and preferential sampling resulting in biases in estimates and predictions, (Chilès, 2012; Kanevski, 2013). In ML, understanding the topology of the input feature space is critical for quantifying its spatial and dimensional resolution, as well as for defining the validity domain (Kanevski, 2013).

Empirically it was shown that cleaning and refining data can result in better model performance compared to just increasing the model complexity and optimization.

Let us note that training multiple models with different origins for the same task yields valuable insights into the data, improves modelling and enriches the interpretation of results.

2 Methodology

Our experience on application of ML to environmental data has resulted in the development of a generic methodology pointing out the most essential phases of the study, figure 1. First, let us precise some important characteristics of environmental data justifying the use of advanced ML techniques.

2.1 Environmental data

Environmental data are an interesting domain of ML application for several reasons, in particular: 1) quantity (small, large and big) and diverse quality; 2) non-linearity; 3) high spatio-temporal variability; 4) dimensionality (often environmental problems are considered in high dimensional feature spaces); 5) noise and uncertainties; 6) presence of extreme values.

Several fundamental problems commonly encountered in environmental data studies can be efficiently addressed by well-developed basic machine learning models, including: clustering – identifying similar groups in data; classification – analysis and prediction of categorical/discrete data; regression – analysis and prediction of continuous data and probability density function modelling and prediction, which plays a central role in environmental risk assessment.

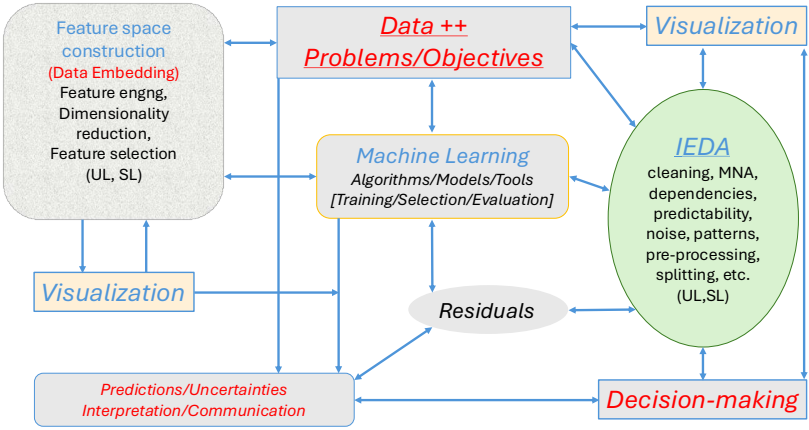


FIGURE 1 – Generic methodology of environmental data ML analysis and modelling. Data++ – raw, reduced and augmented data sets; IEDA – intelligent exploratory data analysis; MNA – monitoring network analysis; UL/SL – unsupervised/supervised learning.

2.2 Data exploration and pre-processing

Exploratory data analysis (EDA) and data pre-processing play a critical role in ML modelling, as the efficacy of data treatment and the results are significantly influenced by the quality and quantity of the input data. They help to better understand the original data and phenomena under study as well as in a proper selection of modelling tools relevant to the objectives of the study. Traditional (EDA) includes techniques to summarize data statistical properties, often using visualization tools, without modelling or making prior hypothesis. It is interesting to note that modern "data science" has its roots in classical EDA (Donoho, 2017).

Contemporary intelligent exploratory data analysis (IEDA) integrates tools and models from statistics, machine learning, and data visualization (Martinez et al., 2022). These tools assist in detecting patterns, quantifying predictability, and constructing the most relevant input space for predictive learning. Efficient and appropriate IEDA is a crucial component of successful data-driven modelling.

Data pre-processing is a key component of IEDA. In addition to

standard techniques such as scaling and normalization, we can highlight several other important methods, including the treatment of missing and extreme values, outlier detection, and feature engineering. Feature engineering involves transforming existing features and creating new ones using domain knowledge to enhance modelling and improve results interpretation. Other processes include data augmentation and splitting the dataset into training, validation, and testing subsets taking into account properties of spatio-temporal data (data clustering, biases, spatial correlations and dependencies) (Kattenborn et al., 2022).

2.3 Data visualization and visual analytics

The importance of visualization in data analysis was well recognized already long time ago by John Tukey (1985): "There is nothing better than a picture for making you think of questions you had forgotten to ask (even mentally)" (Friedman & Stuetzle, 2002, p. 1629). In environmental studies, maps are among the most widely used visualization techniques, effectively summarizing results to support decision-making.

Recently, the field of data visualization has evolved into distinct research domains — visual analytics and visual data mining (Andrienko et al., 2020). Interactive visualization plays a significant role at all stages of environmental data study: from data collection, model construction and learning to the results communication. High dimensional and multivariate data before being visualized usually are processed by applying algorithms and tools for dimensionality reduction, feature selection, projections, etc. (Kuhn & Johnson, 2019; Lee & Verleysen, 2007). Raw high-dimensional data can be visualized and analyzed using popular techniques like parallel coordinates (Inselberg, 2009). Recent advancements in the field of high-dimensional big data visualization have been supported by both algorithmic innovations and hardware improvements, enabling more efficient processing and interpretation of complex datasets.

2.4 Feature selection and dimensionality reduction

The construction of a high-dimensional input feature space relies on three primary sources: 1) expert knowledge of the phenomena and objectives of the study, 2) a critical analysis of existing literature, and

3) feature engineering techniques. However, it is often unclear whether this feature space is comprehensive or contains redundancies. As a result, the input space may include relevant, irrelevant, and redundant features. Therefore, the application of unsupervised and/or supervised feature selection and feature extraction algorithms is essential. These methods not only help reduce dimensionality but also enhance the speed and quality of modeling, while improving interpretability and visualization. (Bolón-Canedo et al., 2015; Guyon & Kacprzyk, 2006; Kuhn & Johnson, 2019; Lee & Verleysen, 2007).

There are three basic approaches to feature selection (FS): 1) *filter methods*, which assess feature relevance independently of machine learning models; 2) *wrapper methods*, which evaluate subsets of features based on model performance; and 3) *embedded methods*, where feature selection is integrated into the training process. The choice of most appropriate FS methods depends on several factors, including data complexity, available computational resources, and the specific model being utilized, since different models may produce varying subsets of features.

Another essential aspect of FS methodology is the concept of intrinsic dimension (ID) of data (Lee & Verleysen, 2007). Numerous methods exist for estimating ID (Camastra & Staiano, 2016). A novel ID estimator based on the multipoint Morisita index introduced initially for spatial data clustering (Morisita, 1959) was proposed in (Golay & Kanevski, 2015). It was demonstrated how Morisita index can be utilized in FS for supervised regression tasks (Golay et al., 2017) and for reducing redundancy in data (Golay & Kanevski, 2017). This research gave an intriguing connection between classical spatial statistics, fractal concepts and contemporary machine learning.

It is important to underline that a well-defined input space in modeling spatial, temporal, or spatio-temporal data can significantly enhance model performance, even when using relatively simple approaches. Conversely, a poorly constructed feature space can lead to suboptimal results, regardless of the complexity or sophistication of the models employed.

3 Machine learning modelling

ML models rely on two types of parameters: hyper-parameters and internal parameters. Hyper-parameters are set by the user before training, while internal parameters are estimated during the training process. For instance, in a multilayer perceptron (MLP), the number of hidden layers and the number of neurons in each layer are hyper-parameters. In contrast, the weights of the connections are computed through optimization algorithms during training.

Fundamentally, learning from data involves two significant steps: model selection and model evaluation. To facilitate these processes, the data are divided into training, validation, and testing subsets. The splitting of geospatial data is a non-trivial task and remains an active area of research (Linnenbrink et al., 2023; Meyer & Pebesma, 2022).

The training subset is used to determine the model's internal parameters, while the validation subset helps identify the optimal combination of hyper-parameters, completing the model selection phase. The test subset is then employed to evaluate the performance of the selected model. The test error provides an estimate of the generalization error, reflecting the model's performance on new unseen data (Bishop & Bishop, 2024; Hastie et al., 2021; Kanevski et al., 2009).

A variety of optimization techniques, e.g., family of gradient descent, are utilized to effectively navigate the solution space and identify optimal or near-optimal parameters for the model. The selection of the cost function is critical, as it influences not only the learning process but also the model's generalization ability to test data.

In more complex scenarios, constraints may be introduced to the optimization problem to account for real-world limitations, for example, expert knowledge, or constraints imposed by physical laws (Karniadakis et al., 2021).

To enhance model performance and prevent overfitting, various effective regularization techniques are also employed (Bishop & Bishop, 2024; Hastie et al., 2021; Haykin, 2009). Regularization helps to constrain model complexity, ensuring that the model not only explains training data but generalizes well.

One of the fundamental questions in data-driven modelling is whether

there exists useful and structured information (i.e., patterns) within the data. Specifically, we seek to determine if the data are predictable within a given input feature space. In the field of geostatistics, this discrimination (pattern – no pattern) can be achieved through the application of variography. When a variogram exhibits a pure nugget effect, it indicates the absence of spatial correlations, signifying the lack of patterns in the data (Chilès, 2012; Kanevski & Maignan, 2004; Kanevsky et al., 1996).

To further assess the presence of patterns, researchers can shuffle the data, resulting in a dataset that maintains the same distribution but destroyed any inherent patterns. Such shuffled datasets serve as control sets, allowing us to evaluate how the model reacts to data lacking of patterns.

A separate and essential question deals with the estimation of the noise in data before modelling. Having this information, we can better perform modelling, avoid overfitting and contribute to the interpretability. There are several non-parametric approaches capable of performing this task (Devroye et al., 2018; Liitiainen et al., 2009).

In summary, data can be represented as: $data = information + noise$ (*unexplained_variability*). The goal of modelling is to extract the information while ensuring that the residuals consist solely of noise. One effective method to achieve this is to shuffle the raw data and the residuals and apply the same analytical approach.

4 Conclusions

Properly applying machine learning in environmental modelling requires deep expertise in both machine learning techniques and the specific data domain. Once the original problem is appropriately reformulated in terms of machine learning, a wide range of models can be employed, including Gaussian processes, random forests, gradient boosting machines, support vector machines, artificial neural networks, deep learning and graph neural networks, among others. These machine learning models are well-developed, optimized, and extensively tested in real-world applications. They are implemented in efficient packages available in R, Python, and other programming languages.

Machine learning is advancing very rapidly, changing fundamentally research domains and practical applications. The integration of openness and reproducibility in modern science makes datasets and codes accessible, which attracts new researchers and accelerates innovation.

To conclude, let us recall some current trends in ML application in environmental studies:

- Nowadays deep learning is an extremely popular approach in environmental applications of machine learning. It drives interest to modern ML and its wide use in applications.

- Physics-informed ML and similar approaches integrate domain expertise and fundamental theories into machine learning models, significantly enhancing their accuracy and reliability in scientific applications.

- Causality. By leveraging ML, researchers have made substantial progress in identifying causal relationships, enabling a deeper understanding of complex systems and improving predictive capabilities (Peters et al., 2017; Runge et al., 2019).

- ML models, in particular deep learning models, are often criticized as "black boxes" due to their complexity and lack of transparency. Recent advancements in Explainable AI (XAI) focus on making these models more interpretable and transparent. XAI aims to bridge the gap between the power of complex ML models and the need for clear, understandable explanations (Molnar, 2018).

- Methodological and practical developments and improvements in data-centric approach are important parts of current innovations in ML.

- Uncertainties quantification and visualization for better environmental risk assessments and intelligent decision making.

And finally, ML is a very useful and stimulating approach in contemporary science worth learning and applying.

Acknowledgements

This paper is dedicated to the memory of Prof. M. Maignan. Collaboration with Michel was always interesting, stimulating and successful. I would also like to thank also to many colleagues and PhD students, for their numerous fruitful discussions on spatial statistics and machine learning, as well as for insights on real data case studies. The improvements in English text were partly made possible thanks to ChatGPT.

References

Akbari, K., Winter, S., & Tomko, M. (2023). Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*, 55(1):56–89.

Andrienko, N., Andrienko, G., Fuchs, G., Slingsby, A., Turkay, C., & Wrobel, S. (2020). *Visual analytics for data scientists*. Springer, Cham, 1st edition.

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., & Mersmann, O. (2023). *Hyperparameter tuning for machine and deep learning with R: A practical guide*. Springer Nature, Singapore.

Bavaud, F. (2009). Information theory, relative entropy and statistics. In Sommaruga, G. (Ed.), *Formal theories of information: From Shannon to semantic information theory and general concepts of information*, pages 54–78. Springer, Berlin, Heidelberg.

Bavaud, F. (2014). Spatial weights: constructing weight-compatible exchange matrices from proximity matrices. In Duckham, M., Pebesma, E., Stewart, K., & Frank, A. U. (Eds.), *Geographic Information Science*, volume 8728, pages 81–96. Springer International Publishing, Cham.

Bavaud, F. (2024). Measuring and testing multivariate spatial autocorrelation in a weighted setting: A kernel approach. *Geographical Analysis*, pages 573–599.

Bishop, C. M. & Bishop, H. (2024). *Deep learning: Foundations and concepts*. Springer International Publishing, Cham.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). *Feature selection for high-dimensional data*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, Cham.

Brus, D. J. (2022). *Spatial sampling with R*. Chapman & Hall/CRC the R Series. CRC Press, Boca Raton, 1st edition.

Camastra, F. & Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41.

Cherkassky, V. S. & Mulier, F. (2007). *Learning from data: Concepts, theory, and methods*. IEEE Press : Wiley-Interscience, Hoboken, NJ, 2nd edition.

Chilès, J.-P. (2012). *Geostatistics: Modeling patial uncertainty*. Number v.713 in Wiley Series in Probability and Statistics Ser. John Wiley & Sons, Incorporated, Hoboken, NJ, 2nd edition.

Devroye, L., Györfi, L., Lugosi, G., & Walk, H. (2018). A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12(1):1752–1778.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

Friedman, J. H. & Stuetzle, W. (2002). John W. Tukey's work on interactive graphics. *The Annals of Statistics*, 30(6):1629–1639.

Gao, B., Wang, J., Stein, A., & Chen, Z. (2022). Causal inference in spatial statistics. *Spatial statistics*, 50:100621.

Gaur, L., Garg, P., & Dey, N. (Eds.) (2023). *Emerging trends, techniques, and applications in geospatial data science*. Advances in Geospatial Technologies. IGI Global, Hershey, PA.

Golay, J. & Kanevski, M. (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index. *Pattern Recognition*, 48(12):4070–4081.

Golay, J. & Kanevski, M. (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension. *Knowledge-Based Systems*, 135:125–134.

Golay, J., Leuenberger, M., & Kanevski, M. (2017). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension. *Pattern Recognition*, 70:126–138.

Guex, G., Loup, R., & Bavaud, F. (2023). Estimation of flow trajectories in a multi-lines transportation network. *Applied Network Science*, 8(1):44.

Guyon, I. & Kacprzyk, J. (2006). *Feature extraction: Foundations and applications*. Number 207 in Studies in Fuzziness and Soft Computing. Springer-Verlag, Berlin.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning*. Springer Texts in Statistics. Springer, New York, NY, 2nd edition.

Haykin, S. S. (2009). *Neural networks and learning machines*. Prentice-Hall, New York, Munich, 3rd edition.

Inselberg, A. (2009). *Parallel coordinates: Visual multidimensional geometry and its applications*. Springer, Dordrecht, NY.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. Springer Texts in Statistics. Springer, New York, NY, 2nd edition.

Jemeljanova, M., Kmoch, A., & Uuemaa, E. (2024). Adapting machine learning for environmental spatial data-a review. *Ecological Informatics*.

Jiang, S., Sweet, L.-b., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., et al. (2024). How interpretable machine learning can benefit process understanding in the geosciences. *Earth's Future*, 12(7):e2024EF004540.

Kanevski, M. (2013). *Advanced mapping of environmental data*. Wiley, Somerset.

Kanevski, M. & Maignan, M. (2004). *Analysis and modelling of spatial environmental data*. EPFL press, Lausanne.

Kanevski, M., Pozdnukhov, A., & Timonin, V. (2009). *Machine learning for spatial environmental data. Theory, applications and software*. EPFL Press.

Kanevsky, M., Arutyunyan, R., Bolshov, L., Demyanov, V., & Maignan, M. (1996). Artificial neural networks and spatial estimation of Chernobyl fallout. *Geoinformatics*, 7(1-2):5–11.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100018.

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. Springer, New York, NY.

Kuhn, M. & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman & Hall/CRC Data Science Series. CRC Press, Taylor & Francis Group, Boca Raton, London, New York, first issued in paperback edition.

Lee, J. A. & Verleysen, M. (Eds.) (2007). *Nonlinear dimensionality reduction*. Information Science and Statistics. Springer, New York, NY.

Liitiainen, E., Verleysen, M., Corona, F., & Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72(16–18):3692–3703.

Linnenbrink, J., Milà, C., Ludwig, M., & Meyer, H. (2023). kNNDM: k-fold nearest neighbour distance matching cross-validation for map accuracy estimation. *EGUsphere*, 2023:1–16.

Mahalle, P. N., Wasatkar, N. N., & Shinde, G. R. (Eds.) (2024). *Data-centric artificial intelligence for multidisciplinary applications*. Chapman & Hall CRC, London.

Martinez, W. L., Martinez, A. R., & Solka, J. (2022). *Exploratory data analysis with MATLAB*. Chapman & Hall CRC, London, 3rd edition.

Meyer, H. & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633.

Meyer, H. & Pebesma, E. (2022). Machine learning based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208.

Molnar, C. (2018). *Interpretable machine learning*. Leanpub, British Columbia, 3rd edition.

Montavon, G., Orr, G. B., & Müller, K.-R. (Eds.) (2012). *Neural networks: tricks of the trade*, volume 7700 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2nd edition.

Morisita, M. (1959). Measuring of dispersion of individuals and analysis of the distributional patterns. *Memoires of the Faculty of Science, Kyushu University, Series E. Biology*, 2:215–235.

Pebesma, E. & Bivand, R. (2023). *Spatial data science: With applications in R*. Chapman & Hall CRC, London.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press, Cambridge, MA, London, England.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. (2019). Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine learning algorithms using spatial data. *Ecological Modelling*, 406:109–120.

Cahiers du CLSL, n° 69, 2025, pp. 143-165

Entre toponymes et situation géographique : cartographie et autocorrélation spatiale des suffixes communaux suisses

Romain Loup
Université de Lausanne
romain.loup@unil.ch

Résumé

Cet article propose une analyse quantitative de la structure des noms de communes en Suisse, en s'intéressant plus particulièrement à la distance toponymiques entre les communes en analysant leur noms et leurs suffixes. En combinant des méthodes d'analyse de noyaux, de mesure de similarité entre noms, et d'autocorrélation spatiale (indice de Moran), il met en évidence une organisation territoriale marquée des suffixes selon les régions linguistiques. L'étude révèle que certains suffixes, comme *-ens*, *-wil* ou *-ano*, présentent une forte concentration géographique, traduisant des dynamiques historiques et culturelles régionales profondes. Cette approche interdisciplinaire, à la croisée de la géographie et des sciences du langage, contribue à une meilleure compréhension de la structuration spatiale des identités toponymiques suisses.

1 Introduction

Les toponymes, bien plus que de simples désignations géographiques, constituent des témoins historiques, linguistiques et culturels des territoires. En Suisse, pays marqué par la coexistence de quatre langues nationales, les noms de communes reflètent une diversité phonétique et morphologique exceptionnelle. Chaque suffixe, chaque terminaison, porte la trace d'une stratification historique profonde, témoignant des migrations, des colonisations et des évolutions linguistiques successives.

La tradition philologique suisse a largement documenté cette richesse. Gatschet (1867) identifie des couches celtiques, romaines,

germaniques et romanes dans l'onomastique nationale. Kristol (2023) dresse un historique qui montre que dès les années 1960, des historiens approfondissent cette analyse en mettant en évidence la persistance régionale de certains suffixes, tels que *-ens* en Suisse romande, *-wil* et *-dorf* en Suisse alémanique, ou *-ano* au Tessin. Plus récemment, des synthèses comme celle de Jordan (2012) ou Cattin et al. (2005) rappellent que les toponymes participent à la construction d'identités spatiales et culturelles.

> Étudier les noms de lieux, c'est découvrir à quel point les différentes régions linguistiques de la Suisse sont mutuellement imbriquées, et interconnectées avec les régions européennes voisines (Cattin et al., 2005, p. 9)

Cependant, malgré l'abondance d'études descriptives, peu de recherches ont quantifié la structure phonétique et morphologique des toponymes et leur organisation spatiale. Les avancées en statistique spatiale – notamment l'application des indices d'autocorrélation, du *Moran scatterplot* (Anselin, 1996) et des techniques de projection multidimensionnelle (MDS) (Borg & Groenen, 1997) – offrent de nouveaux outils pour explorer ces questions de manière rigoureuse (cf. Bavaud, 2014, 2023). Le présent article propose ainsi une approche quantitative originale de la toponymie suisse, en s'appuyant sur :

— l'extraction systématique des suffixes communaux,

— la construction d'une distance toponymique entre les communes,

— l'analyse de l'organisation spatiale de ces structures onomastiques.

En analysant la structure de chaque nom de commune ainsi que leur suffixe, nous visons à explorer deux axes principaux. (I) Les suffixes des noms de communes présentent-ils des structures régionales marquées en Suisse ? (II) Pouvons-nous, à partir d'une analyse phonétique ou onomastique, retrouver les dynamiques linguistiques et historiques connues du territoire ? Ainsi, en combinant méthodes de phonétique computationnelle, statistique spatiale, et cartographie analytique, cette étude souhaite contribuer à la compréhension de la géographie linguistique de la Suisse contemporaine.

2 Données et préparation

L'analyse repose sur l'ensemble des 2126 communes suisses recensées en 2024, couvrant les quatre principales aires linguistiques : la Suisse alémanique, la Suisse romande, Suisse italienne et la Suisse romanche. Les données utilisées (Office Fédéral de la Statistique, 2024) combinent des informations toponymiques, linguistiques et géographiques :

— noms officiels des communes et des localités,
— coordonnées géographiques (longitude, latitude) des centres communaux,
— langue principale de chaque commune (classée en quatre catégories : allemand, français, italien, romanche).

La préparation des données détaillées ci-après a suivi plusieurs étapes, visant à standardiser les formes linguistiques et à permettre une analyse toponymique robuste.

2.1 Nettoyage des noms

Les noms de communes ont été systématiquement nettoyés et standardisés afin d'homogénéiser leur représentation écrite et simplifier le traitement informatique :

— suppression des accents et caractères spéciaux (*é*, *ä*, *ç*, etc.),
— conversion en minuscules,
— suppression des éléments non alphabétiques (parenthèses, traits d'union, apostrophes).

2.2 Extraction des suffixes

Afin d'analyser la structure morphophonologique des noms, des suffixes de longueur variable ont été extraits :

— Suffixes de 3 lettres (par ex. *-ens*, *-wil*, *-(d)orf*, *-ano*),
— avec en complément des tests sur des suffixes de 1 à 4 lettres pour explorer d'autres régularités.

Chaque commune a ainsi été associée à un suffixe principal, défini comme les dernières lettres significatives du nom nettoyé.

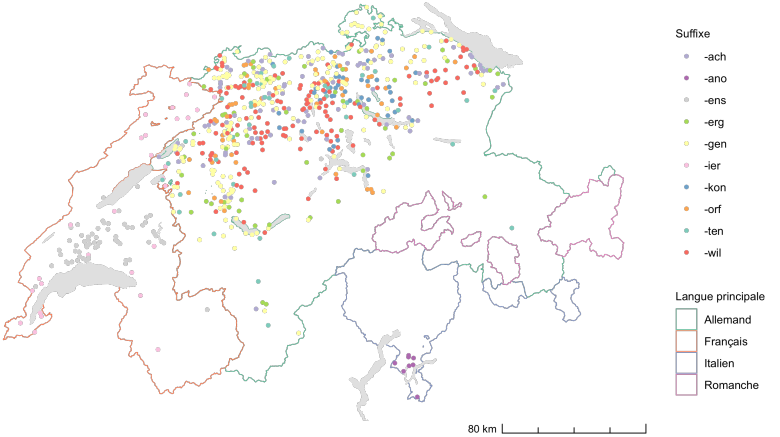


FIGURE 1 – Suffixes de trois lettres les plus fréquents en allemand, français et italien.

2.3 Que signifient ces suffixes ?

Les suffixes toponymiques suisses portent la trace de strates historiques et linguistiques profondes. Ils permettent de retracer les mouvements de peuplement, les dynamiques culturelles et les influences croisées entre langues germaniques et romanes. Plusieurs familles de suffixes sont particulièrement révélatrices.

Parmi les plus répandus en Suisse occidentale figurent les formes issues du suffixe germanique *-ingōs*, devenu *-ens*, *-in* ou *-an*, souvent combiné à un anthroponyme germanique. Très présent dans les cantons de Vaud et de Fribourg, ce suffixe indique l'appartenance à un clan ou un groupe familial : il signifie littéralement « chez les gens de », le premier élément désignant le fondateur de la localité (Cattin et al., 2005, p. 117). Une variante, *-ingun* (> *-ingen*, *-igen*), prolonge cette logique d'identification collective au travers de la notion de suite ou de clan (Cattin et al., 2005, p. 314).

Les formes romanes de suffixes dérivées du latin *-acum* ou *-akos*

donnent naissance à deux séries distinctes selon la région linguistique. En Suisse romande, ces formes aboutissent le plus souvent à *-y*, *-ier*, *-ey*, *-ay* ou *-ex* (comme dans Giez VD, Delley FR, Cossonay VD ou Thônex GE). Ces suffixes proviennent de gentilices ou de cognomina romains et ont progressivement évolué en fonction des dynamiques phonétiques locales. En Suisse alémanique, en revanche, la même racine latine se fossilise dans la forme germanisée *-ach* (Dornach SO, Bülach ZH), tandis qu'au Tessin elle devient *-ago* (Brissago TI) (Cattin et al., 2005, p. 208).

Autre exemple d'adaptation germanique, le suffixe *-(i)kofen* ou *-(i)kon*, dérivé de *-ing -hoven*, désigne des habitats aménagés, apparus entre les VII^{e} et XVIII^{e} siècles. On le retrouve sous la forme abrégée dans les cantons d'Argovie, de Lucerne et de Zurich, et sous la forme longue dans le canton de Thurgovie ou à l'ouest de la Suisse alémanique (notamment Berne et Soleure) (Cattin et al., 2005, p. 990).

Le cas du suffixe *-wil* est également emblématique. Il remonte au vieux haut allemand *wilāri*, issu du roman *villare*, lui-même formé sur l'adjectif latin *villaris* (dérivé de *villa*, désignant un domaine agricole). Ce suffixe renvoie à un petit village ou un hameau isolé. On le retrouve aussi bien en Suisse alémanique (Wil SG, Horgen ZH) que sous ses équivalents romans dans des noms comme Villars FR ou Villiers NE (Cattin et al., 2005, p. 673).

Enfin, dans les régions italophones, on trouve des formes en *-ano* ou *-an*, issues du latin *-anum*, *-ana*, *-ano*, qui servent à former des adjectifs de relation ou de possession (p. ex. proche de, appartenant à) (Cattin et al., 2005, p. 549 ; Vassere, 1996, p. 1443).

Il est à noter que ces suffixes restent en grande partie spécifiques à chaque région linguistique. On observe très peu de chevauchements entre les aires germanophone, francophone et italophone, ce qui renforce l'idée d'une forte structuration linguistique et culturelle des formes toponymiques, comme le montre la carte de la figure 1.

2.4 Construction de la distance toponymique

Deux approches principales ont été envisagées pour quantifier la similarité phonétique entre les noms de communes. La première, divisée

en deux, repose sur une comparaison directe des chaînes de caractères, soit par la **distance de Levenshtein** sur les noms nettoyés (Levenshtein, 1966), soit après **phonétisation** des noms à l'aide d'algorithmes comme Soundex ou NYSIIS (Odell & Strong, 1947). La seconde approche, plus structurelle, consiste à représenter chaque nom de commune sous forme d'un **vecteur de bigrammes** (séquences de deux lettres successives) et à mesurer leur proximité à l'aide d'une **distance cosinus** (Cohen, 2000; Cohen et al., 2003).

Après évaluation, la distance cosinus sur les vecteurs de bigrammes a été privilégiée car elle est indépendante des langues, elle capture efficacement les structures syllabiques et morphologiques typiques des toponymes suisses et est robuste aux variations orthographiques locales. Le résultat est une matrice de distances toponymique $\mathbf{D}_{\text{nom}}$ symétrique, de dimension 2126×2126.

2.5 Projections et voisinages spatiaux

En parallèle des distances toponymiques, deux structures de voisinage ont été construites : voisinage géographique basé sur les 10 plus proches voisins en coordonnées XY (k-NN), et le voisinage toponymique basé sur les 10 plus proches voisins dans l'espace MDS toponymique.

Ces structures permettent d'analyser la cohérence spatiale et toponymique des suffixes, via des mesures d'autocorrélation globale et locale (*I* de Moran).

3 Méthodes

L'objectif méthodologique est de quantifier la similarité toponymique entre les noms de communes suisses et d'analyser leur structuration spatiale. Les trois approches principales sont détaillées par la suite.

3.1 Distance de Levenshtein sur les noms nettoyés

La distance de **Levenshtein** d_L mesure le nombre minimal d'opérations élémentaires (insertion, suppression ou substitution de caractères) nécessaires pour transformer un mot en un autre (Levenshtein, 1966). Elle est définie par l'équation de récurrence :

$$d_L(i,j) = \begin{cases} d_L(0,0) = 0 \\ \min \begin{cases} d_L(i-1,j)+1 \\ d_L(i,j-1)+1 \\ d_L(i-1,j-1) + \mathbf{1}_{s_i \neq t_j} \end{cases} & \text{sinon} \end{cases} \quad (1)$$

et la distance entre s et t est finalement $d_L(|s|,|t|)$.

Cette mesure – directement appliquée aux noms nettoyés – est sensible aux variations locales d'orthographe, mais peut être influencée par la longueur des noms.

3.2 Phonétisation suivie de comparaison

Pour atténuer les effets d'orthographe et se rapprocher de la réalité phonétique, les noms de communes ont été transformés à l'aide d'algorithmes de phonétisation : **Soundex**, conçu à l'origine pour l'anglais (Odell & Strong, 1947) et **NYSIIS**, légèrement mieux adapté aux langues romanes, car il groupe plusieurs lettres ensemble (Anon, 1968).

Chaque nom s est transformé en une représentation phonétique $\phi(s)$. La distance phonétique devient alors la distance de Levenshtein entre les représentations phonétiques :

$$d_\phi(s,t) = d_L(\phi(s), \phi(t)) \quad (2)$$

Cette approche standardise les formes similaires (p. ex. *Lausanne vs Lozan*) mais reste dépendante du choix de l'algorithme de phonétisation.

De plus, cette distance est sensible à la langue d'origine du toponyme : les algorithmes de phonétisation comme **Soundex** ou **NYSIIS** ont été conçus pour l'anglais et peuvent mal représenter les particularités phonologiques du français, de l'allemand ou de l'italien (Christian, 1998).

3.3 Distance cosinus sur vecteurs de bigrammes

Enfin, une approche plus structurelle a été utilisée. Elle consiste à représenter chaque nom de commune comme un **vecteur de bigrammes** de caractères, c'est-à-dire toutes les séquences de deux lettres successives apparaissant dans le mot. Cette représentation permet de capturer la forme morphographique des toponymes, en tenant compte des enchaînements de lettres typiques des langues concernées. L'approche est largement utilisée en traitement automatique des noms propres, en particulier dans les tâches de désambiguïsation ou de rapprochement d'entités textuelles similaires (Cohen et al., 2003). Plus formellement, à chaque mot s est associé un vecteur $v(s) \in \mathbb{R}^d$, où chaque composante correspond à la fréquence (ou la présence) d'un bigramme parmi les d bigrammes possibles (par exemple, *su*, *ul*, *ll*, *le*, *en*, etc.). La similarité entre deux noms s et t est alors mesurée par la **similarité cosinus** :

$$\cos(s,t) = \frac{v(s) \cdot v(t)}{\|v(s)\|\|v(t)\|} \tag{3}$$

Soient $v(s)$ et $v(t)$ les vecteurs de fréquences de bigrammes associés aux noms s et t. La **distance phonétique cosinus** dérivée est :

$$d_{\cos}(s,t) = 1 - \cos(s,t) \tag{4}$$

Cette approche est largement utilisée en traitement automatique des langues (TAL) pour mesurer la similarité entre mots, noms propres ou entités nommées, en particulier dans les contextes multilingues ou bruités où les distances basées sur l'orthographe stricte (comme Levenshtein) sont trop sensibles.

Les bigrammes permettent de modéliser de façon souple les motifs morphologiques (préfixes, suffixes, syllabes) sans recourir à une phonétisation complexe. Leur robustesse a été démontrée dans des tâches de désambiguïsation de noms propres, de reconnaissance d'entités nommées ou de recherche floue (Cohen et al., 2003; Kondrak, 2005). Cette mesure présente plusieurs avantages : elle est indépendante de la langue et de l'orthographe précise, elle capte les similarités

morphologiques globales (préfixes, suffixes, rythmes syllabiques) et elle est robuste aux différences mineures de longueur et d'orthographe entre noms.

Ces propriétés en font un outil particulièrement bien adapté à la diversité linguistique suisse, où les suffixes sont des marqueurs à la fois phonétiques, morphologiques et régionaux. Il a par conséquent été retenu pour la suite de l'analyse.

3.4 Construction de la matrice de distances toponymiques et vecteur de la distance locale

À partir de la méthode choisie, une **matrice de distances toponymiques** $\mathbf{D}_{\text{nom}}$ de dimension 2126×2126 a été construite, symétrique, carrée et de diagonale nulle. Cette matrice constitue le support principal pour les analyses de structuration spatiale et phonétique présentées dans la suite de l'article. Le vecteur de pondération $\mathbf{f}$ où $f_i = n_i/n_\bullet$ représente le poids spatial et n_i correspond au nombre d'habitants dans une commune i. Afin de quantifier l'hétérogénéité toponymique du voisinage immédiat de chaque commune, nous avons construit un indicateur de dissimilarité toponymique locale. Pour chaque commune i, nous avons identifié ses dix voisines géographiques les plus proches (en distance à vol d'oiseau), et calculé la moyenne des distances toponymique dl_i entre cette commune et ses voisines. Le vecteur de la **distance locale** dl_i repose sur une mesure préétablie de dissimilarité toponymique entre les noms des communes. Cette moyenne locale fournit un score de dissemblance contextuelle qui reflète dans quelle mesure une commune s'écarte phonétiquement de son environnement spatial immédiat. Cet indicateur constitue la variable d'entrée de l'analyse d'autocorrélation spatiale présentée ci-après (sec. 4.2.3),

$$\mathbf{dl}_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{D}_{ij}^{\text{nom}} \tag{5}$$

où $\mathcal{N}_i$ est l'ensemble des $k = 10$ plus proches voisines géographiques de la commune i, et $\mathbf{D}_{ij}^{\text{nom}}$ désigne la distance toponymique entre les communes i et j.

4 Résultats : toponymie dans les communes

L'analyse s'appuie sur la matrice de distances toponymiques construite à partir des vecteurs de bigrammes des noms de communes. Les résultats sont présentés en trois volets : la structure toponymique générale, la distribution spatiale des suffixes et l'autocorrélation locale des formes toponymiques. Les analyses seront menées à la fois avec pondération $\mathbf{f}$, selon le poids de communes, et avec des poids uniformes ($\mathbf{f}_{\text{uni}} = 1/n$). Cette double approche permet de distinguer deux dimensions complémentaires : d'une part, la structuration toponymique intrinsèque des noms, et d'autre part, l'influence potentielle de la taille ou du poids démographique des communes.

4.1 Structure toponymique générale

À partir de la matrice symétrique de distances toponymique $\mathbf{D}_{\text{nom}}$, des **analyses multidimensionnelles classiques et pondérées (MDS)** ont été réalisées. La projection sur les deux premiers axes permet de visualiser la structuration toponymique globale.

Le MDS pondéré (ou de poids uniformes) utilisé est basé sur la matrice de centrage $\mathbf{H}$ et le noyau de caractéristiques (*feature kernel*) $\mathbf{K}_X$ (Bavaud, 2024), définis par :

$$\mathbf{H} = \mathbf{I}_n - \mathbf{1}_n \mathbf{f}^\top \in \mathbb{R}^{n \times n} \quad \text{et} \tag{6}$$

$$\mathbf{K_X} = -\frac{1}{2} \mathbf{\Pi}^{\frac{1}{2}} \mathbf{H} \mathbf{D}^2 \mathbf{H}^\top \mathbf{\Pi}^{\frac{1}{2}} \tag{7}$$

où $\mathbf{f}$ est le vecteur des poids relatifs des communes, $\mathbf{1}$ est le vecteur colonne de taille n contenant des 1 et $\mathbf{\Pi} = \text{diag}(\mathbf{f}) \in \mathbb{R}^{n \times n}$ est la matrice diagonale de poids.

La projection toponymique finale $\mathbf{X}$ du MDS (Bavaud, 2023) est obtenue par :

$$\mathbf{X} = \mathbf{\Pi}^{-1/2} \mathbf{U} \, \Lambda^{1/2}, \quad x_{i\alpha} = \frac{1}{\sqrt{f_i}} u_{i\alpha} \sqrt{\lambda_\alpha} \tag{8}$$

où $\mathbf{U}$ et Λ sont respectivement les vecteurs et valeurs propres de $\mathbf{K_X}$.

4.2 Tester et visualiser l'autocorrélation

Cette partie a pour but de comprendre dans quelle mesure les noms des communes sont autocorrélés, s'il y a des groupe de noms qui se forment et comment la pondération $\mathbf{f}$ joue un rôle dans les représentations graphiques. Les scores seront aussi comparés avec d'autres indicateurs.

4.2.1 Scores d'autocorrélation spatiale

Comme défini par Loup & Bavaud (2025, p. 5, Eq. 7), le test

$$z = \frac{\delta - \mathbb{E}_0(\delta)}{\sqrt{\mathbb{V}\mathrm{ar}(\delta)}} > u_{1-\alpha} \tag{9}$$

où $u_{1-\alpha}$ représente le quantile standard normal permet de rejeter l'hypothèse H_0 au niveau α. Le tableau 1 présente les **valeurs standardisées des scores** z, mesurant l'autocorrélation spatiale entre le noyau toponymique $\mathbf{K}_{\text{nom}}$ (calculé à partir d'une distance cosinus sur les bigrammes) et six noyaux d'interprétation spatiale : politique ($\mathbb{K}_{\text{pol}}$), temps de trajet ($\mathbb{K}_{\text{time}}$), taille des communes ($\mathbb{K}_{\text{size}}$), langue ($\mathbb{K}_{\text{ling}}$), richesse ($\mathbb{K}_{\text{rich}}$), et transport optimal des richesses ($\mathbb{K}_{\text{TO}}$). On observe que l'autocorrélation est particulièrement importante pour plusieurs de ces dimensions, notamment pour le noyau politique, mais aussi pour celui reflétant la taille des communes. Ces valeurs sont bien au-dessus de scores usuels d'une loi normale. Il est remarquable que les noyaux politique et linguistique présentent une autocorrélation plus forte que le noyau fondé sur le temps de trajet, suggérant que les clivages symboliques et institutionnels structurent l'espace aussi fortement, voire davantage, que les contraintes de mobilité. Cela appuie l'idée que les noms de lieux en Suisse portent la trace d'un enracinement local durable, en lien avec les dynamiques territoriales historiques, culturelles et politiques. En revanche, les noyaux liés à la richesse et au transport optimal montrent des niveaux d'autocorrélation plus faibles, traduisant une relation moins directe entre les toponymes et les dimensions socio-économiques contemporaines.

	$\mathbb{K}_{\text{pol}}$	$\mathbb{K}_{\text{time}}$	$\mathbb{K}_{\text{size}}$	$\mathbb{K}_{\text{ling}}$	$\mathbb{K}_{\text{rich}}$	$\mathbb{K}_{\text{TO}}$
$\mathbf{K}_{\text{nom}}$	200.8	118.3	189.4	132.7	40.6	95.1

TABLEAU 1 – Scores d'autocorrélation spatiale standardisés (z) mesurant l'alignement entre le noyau toponymique $\mathbb{K}_{\text{nom}}$ (ligne) et six noyaux spatiaux (colonnes) : politique, temps de trajet, taille des communes, langue, richesse, et transport optimal des richesses.

4.2.2 MDS : Des groupements bien marqués

La **carte toponymique** obtenue révèle plusieurs regroupements remarquables : les communes de Suisse romande se rassemblent autour de suffixes en *-ens*, *-ier*, *-res*, *-gny* alors que les communes alémaniques forment des ensembles distincts autour de *-gen*, *-wil*, *-ach*, *-(d)orf*. Enfin les communes tessinoises et grisonnes se regroupent dans une zone associée aux suffixes en *-ano*, *-one*.

Les figures 2 et 3 présentent les deux premiers axes de la décomposition multidimensionnelle classique (MDS), respectivement dans les versions pondérée et non pondérée. Dans les deux cas, on observe une structure en *clusters* bien marquée, où certains suffixes apparaissent fortement regroupés : c'est notamment le cas pour les suffixes *-gen*, *-wil*, *-ach*, *-ens* et *-ano*. La pondération par la taille des communes (Fig. 2, *scree graph*, annexes Fig. 7) a pour effet d'étaler la projection et d'éloigner certaines communes plus influentes démographiquement. Cela rend visibles des effets de masse : les communes de grande taille occupent des positions plus centrales ou plus isolées selon leur profil toponymique. La figure 3 – pondération uniforme – regroupe encore plus les suffixes et les effets de taille disparaissent.

On note par exemple que la ville de Berne est projetée légèrement à l'écart du centre de gravité des autres points, ce qui pourrait s'expliquer par la particularité phonétique de son suffixe et par son profil linguistique hybride ou neutre dans un espace fortement structuré par les oppositions linguistiques. En effet, *Berne* est certainement un mot celtique donné par le nom d'une rivière (Cattin et al., 2005, p. 143 ; Müller, 2002, p. 87). Cette hypothèse est renforcée par la présence de regroupements linguistiques nets dans les deux configurations : les

suffixes germanophones, francophones et italophones tendent à occuper des zones distinctes dans l'espace factoriel. Il est particulièrement intéressant de constater que cette structuration linguistique est aussi marquée que celle induite par les suffixes eux-mêmes, illustrant la profondeur culturelle et territoriale des régularités toponymiques en Suisse.

FIGURE 2 – MDS sur les toponymes, pondéré par le poids des communes et les 50 plus grandes communes nommées.

4.2.3 Autocorrélation spatiale des dissimilarités toponymiques : effets d'échelle et de pondération

Afin d'évaluer dans quelle mesure les dissimilarités toponymiques locales entre les communes suisses présentent une organisation spatiale, nous avons mobilisé l'indice de Moran pondéré (Bavaud, 2024). Cette

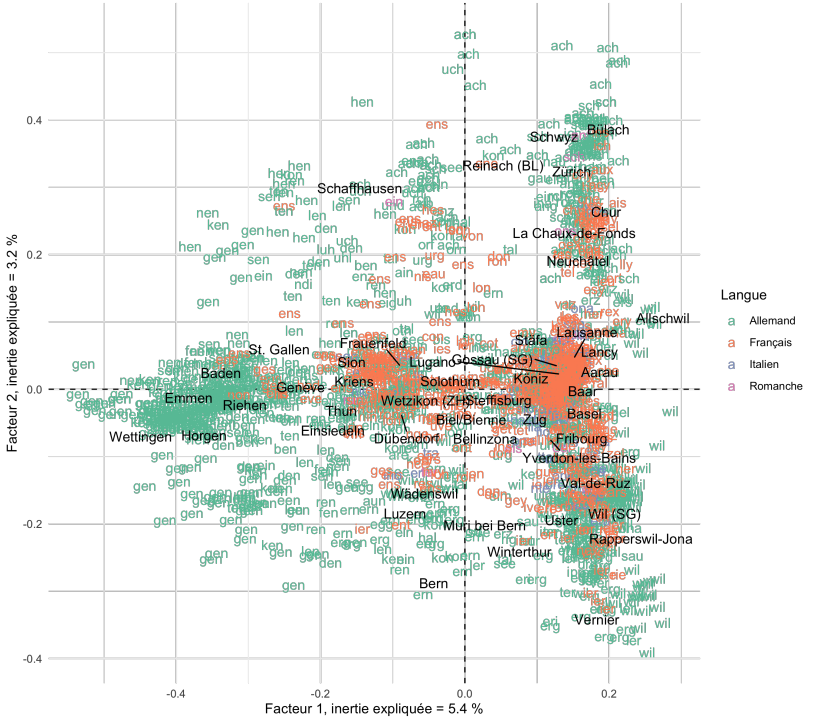


FIGURE 3 – MDS sur les toponymes, pondération uniforme et les 50 plus grandes communes nommées.

mesure, adaptée à nos données par l'intégration d'un noyau de chaleur (*heat kernel*), permet d'apprécier l'autocorrélation spatiale à différentes échelles, modulées par le paramètre de diffusion t (Loup & Bavaud, 2025, Eq. 23, p. 13). Deux types de pondération ont été testés : une pondération uniforme (chaque commune ayant le même poids dans l'ensemble) et une pondération régionale, fondée sur un vecteur $\mathbf{f}$ reflétant l'importance démographique des communes.

Les *scatterplots* de Moran-Anselin (Fig. 4) obtenus pour une faible valeur de $t = 1$, donc pour une interaction spatiale fortement locale, révèlent une autocorrélation toponymique marquée, avec une pente de 0.66 dans le cas d'une pondération uniforme, et de 0.5 sous pon-

dération régionale. Ces résultats indiquent que, localement, les communes voisines ont tendance à présenter des structures toponymiques similaires, suggérant une certaine continuité onomastique à l'échelle micro-régionale.

Lorsque le paramètre t est augmenté à 50 (Fig. 5), étendant l'influence spatiale à des voisinages beaucoup plus larges, l'intensité de l'autocorrélation diminue significativement et des groupes de noms romands et alémaniques se séparent verticalement. La pente du *scatterplot* tombe alors à 0.13 (pondération uniforme) et 0.07 (pondération régionale), révélant une structure spatiale plus diffuse et une perte de cohérence phonétique à large échelle. Autrement dit, les similitudes phonétiques observées à petite échelle tendent à s'estomper lorsqu'on considère des relations intercommunales plus distantes, ce qui atteste du caractère principalement local des régularités onomastiques.

L'effet de la pondération est également instructif. Dans tous les cas, la pondération régionale ($\mathbf{f}$) tend à réduire l'estimation de l'autocorrélation par rapport à la pondération uniforme. Cela suggère que les grandes communes – qui reçoivent un poids plus important – affichent une dissimilarité toponymique plus marquée par rapport à leur voisinage immédiat. Ce phénomène peut être interprété comme le reflet d'une plus grande hétérogénéité interne ou d'une situation linguistique plus complexe dans les centres urbains, souvent situés aux interfaces entre régions linguistiques ou dotés d'une toponymie historiquement composite. De plus, la variance expliquées de deux premiers scores factoriels est un peu plus élévée dans ce cas.

Ces résultats plaident ainsi pour une lecture multi-échelle de l'organisation spatiale des noms de communes, dans laquelle la cohérence phonétique apparaît fortement structurée à l'échelle locale, mais perd en intensité lorsque l'on s'éloigne géographiquement ou lorsque l'on pondère par des poids démographiques. L'approche par noyau de chaleur permet de moduler finement cette échelle d'analyse et de mieux cerner les dynamiques linguistiques et historiques qui sous-tendent la géographie des toponymes suisses.

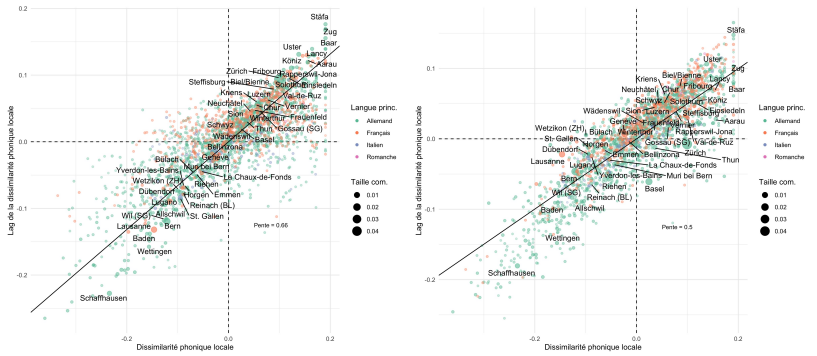


FIGURE 4 – Scatterplots de Moran-Anselin, $t = 1$. Gauche : poids $\mathbf{f}$. Droite : poids uniformes. 50 plus grandes communes nommées.

4.3 Distribution spatiale des suffixes

Pour approfondir la structuration régionale, les suffixes de trois lettres les plus fréquents ont été sélectionnés. Des cartes ont été réalisées pour visualiser leur répartition (Fig. 1 et 8) ainsi qu'un tableau (2) regroupant les suffixes les plus fréquents avec leur nombre d'occurrences.

Ces cartes mettent en évidence une forte concentration des suffixes *-ens* dans les cantons de Vaud et de Fribourg, la prédominance des suffixes *-wil* et *-gen* dans la Suisse orientale, la localisation des suffixes *-ano*, *-one* exclusivement dans le Tessin. La structure phonétique des suffixes suit donc clairement la répartition historique des langues nationales et de leurs variantes dialectales.

Allemand (1373)	**Français** (615)	**Italien** (122)
-gen (168)	*-ens* (61)	*-ano* (8)
-wil (134)	*-ier* (25)	*-one* (7)
-ach (84)	*-res* (18)	*-gio* (4)
-orf (55)	*-gny* (17)	*-gno* (4)
-eng (54)	*-lle* (17)	*-ino* (4)

TABLEAU 2 – Occurrences des suffixes les plus fréquents pour les trois langues principale.

FIGURE 5 – Scatterplots de Moran-Anselin, $t = 50$. Gauche : poids $\mathbf{f}$. Droite : poids uniformes. 50 plus grandes communes nommées.

Afin de comprendre de façon quantitative ces observations, le tableau 3 présente les indices de Moran (non pondérés) calculés pour une sélection de suffixes toponymiques parmi les plus fréquents en Suisse, ainsi que leur langue d'attache et la significativité statistique de leur distribution spatiale. L'indice de Moran I mesure ici le degré d'autocorrélation spatiale d'un suffixe donné : une valeur élevée indique que les communes partageant ce suffixe ont tendance à être spatialement proches les unes des autres, signalant une concentration régionale.

Suffixe	**Langue**	**Moran I**	**Valeur p**
-ens	Fra	0.23	1.5×10^{-135}
-ins	Fra	0.16	1.8×10^{-73}
-wil	All	0.15	1.7×10^{-63}
-gen	All	0.14	4.1×10^{-49}
-kon	All	0.10	3.1×10^{-30}
-see	All	0.07	6.2×10^{-16}
-orf	All	0.07	2.9×10^{-14}
-ach	All	0.06	5.8×10^{-13}
-ges	All	0.06	1.6×10^{-13}
-ken	All	0.05	1.2×10^{-9}

TABLEAU 3 – Dix indices de Moran les plus forts pour des suffixes communaux selon la langue.

Comme le montre le tableau 3, le suffixe *-ens*, particulièrement romand, présente la plus forte autocorrélation ($I = 0.23$, $p < 10^{-130}$), suivi par *-ins*, également en français. « En Suisse occidentale, il existe trois régions qui connaissent une densité significative de noms de lieux en *-ānum* (<*ins*) : la région de Nyon, colonie romaine, celle d'Avenches, ancienne capitale de l'Helvétie romaine, et celle de Studen (la Petinesca romaine, au bout du lac de Bienne) » (Cattin et al., 2005, p. 715). Cette forte structuration spatiale suggère que certains suffixes romands sont fortement enracinés dans des zones géographiques précises, vraisemblablement en lien avec l'histoire linguistique et la toponymie locale de la Suisse romande (cf. annexes, Fig. 8.

Du côté alémanique, plusieurs suffixes présentent également une autocorrélation marquée, tels que *-wil* ($I = 0.15$), *-gen* ($I = 0.14$) ou *-kon* ($I = 0.10$) (Fig. 6). Ces résultats indiquent que certains éléments lexicaux propres au suisse allemand ont également une répartition géographiquement structurée, même si l'intensité de leur autocorrélation est légèrement inférieure à celle observée pour les suffixes romands.

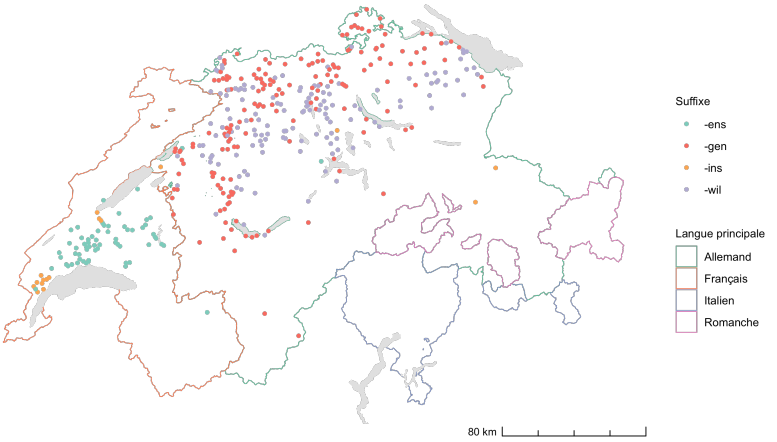


FIGURE 6 – Les quatre suffixes les plus autocorrélés localisés.

L'ensemble des suffixes retenus présentent une significativité statistique très forte ($p < 10^{-8}$), ce qui confirme que la distribution spatiale des éléments phonétiques des noms de communes en Suisse n'est pas aléatoire, même si les scores du I restent contenus. Elle témoigne au contraire de dynamiques régionales cohérentes, reflétant à la fois les frontières linguistiques, les histoires de peuplement et les influences morphologiques propres à chaque aire culturelle.

4.4 De commune à localité

À partir des 2126 communes suisses, les mêmes analyses peuvent être étendues aux 5764 localités définies par l'Office Fédéral de la Topographie (2024), dont 3973 portent un nom unique. L'analyse à cette échelle plus fine renforce les résultats obtenus : pour le test du I de Moran, les suffixes *-wil* (0.24), *-ens* (0.24), *-see* (0.20) ou *-gen* (0.20) présentent des valeurs d'autocorrélation encore plus élevées, accompagnées de p-valeurs encore plus significatives que celles observées au niveau communal. Cela s'explique en partie par le fait qu'en Suisse alémanique – plus encore qu'en Romandie – les fusions de localités en communes ont souvent atténué les effets de proximité entre noms partageant des suffixes similaires.

5 Conclusion

Cette contribution visait à mettre en avant l'interdisciplinarité de travaux à la croisée de la géographie et des lettres, en mêlant la distance spatiale entre des communes et leurs toponymes. Comme nous avons pu le voir tout au long de cet article, la géographie se mêle à l'histoire et à la formation des noms. Cela est mis en évidence grâce à plusieurs méthodes quantitatives, telles que la transformation de distances en noyaux qui a permis une visualisation en MDS montrant déjà une spatialisation claire des noms. L'autocorrélation de toponyme a aussi pu être vérifiée de plusieurs manières, par l'étude du noyau, comparé à plusieurs autres distances, mais aussi à l'échelle des suffixes. L'analyse de la répartition des toponymes dans les noms de communes suisses ainsi que de leurs suffixes révèlent une organisation spatiale loin d'être aléatoire. Plusieurs suffixes – notamment *-ens*, *-wil*,

-gen ou *-ach* – présentent une autocorrélation significative, indiquant une concentration régionale nette. Cette structuration n'est pas homogène : elle est plus marquée pour les suffixes francophones que pour les suffixes alémaniques, ce qui suggère des dynamiques historiques et linguistiques distinctes selon les aires culturelles. Les résultats obtenus à l'aide de l'indice de Moran pondéré montrent que l'autocorrélation toponymique est avant tout locale. Les régularités s'estompent rapidement à mesure que l'on élargit l'échelle spatiale d'interaction, comme en témoignent les courbes décroissantes obtenues avec le noyau de chaleur. Ce comportement met en lumière le rôle structurant des micro-régions linguistiques, souvent invisibles dans les découpages administratifs, mais bien ancrées dans l'onomastique. L'introduction d'un vecteur de pondération régional dans le calcul de l'indice de Moran nuance encore cette lecture : les grandes communes ou celles ayant un poids démographique plus important tendent à présenter une dissimilarité toponymique plus élevée par rapport à leur environnement immédiat. Ce phénomène pourrait s'interpréter comme le reflet d'une plus grande hétérogénéité interne, ou d'une position liminaire à l'interface entre espaces linguistiques.

Ces résultats s'inscrivent dans le sillage des travaux de François Bavaud sur la formalisation de la proximité spatiale et la modélisation des structures morpho-linguistiques régionales. L'analyse toponymique proposée ici rejoint, dans une perspective morphogénétique, des approches similaires appliquées au champs politique ou économique. Les suffixes toponymiques, au-delà de leur portée linguistique, apparaissent ainsi comme des marqueurs spatiaux d'appartenances régionales cohérentes, et constituent une voie féconde pour étudier la formation, la persistance et la diffusion des identités territoriales.

Références

Anon (1968). *NYSIIS (New York State Identification and Intelligence System)*. Office of Justice Programs, New York, NY, 2ème édition.

Anselin, L. (1996). The Moran scatterplot as an ESDA Tool to assess local instability in spatial association. In Fischer, M. (éd.), *Spatial Analytical*,

pages 111–125. Taylor and Francis, an imprint of Routledge, Boca Raton, FL, 1ère édition.

Bavaud, F. (2014). Spatial weights: Constructing weight-compatible exchange matrices from proximity matrices. In Duckham, M., Pebesma, E., Stewart, K., & Frank, A. U. (éd.), *Geographic Information Science*, volume 8728, pages 81–96. Springer International Publishing, Cham.

Bavaud, F. (2023). Exact first moments of the RV coefficient by invariant orthogonal integration. *Journal of Multivariate Analysis*, 198:105227.

Bavaud, F. (2024). Measuring and testing multivariate spatial autocorrelation in a weighted setting: A kernel approach. *Geographical Analysis*, 56(3):573–599.

Borg, I. & Groenen, P. (1997). *Modern multidimensional scaling: theory and applications*. Springer, New York, NY, USA, 1ère édition.

Cattin, F., Aquino-Weber, D., Kristol, A. M., & Université de Neuchâtel (éd.) (2005). *Dictionnaire toponymique des communes suisses : DTS = Lexikon der schweizerischen Gemeindenamen*. Huber [u.a.], Frauenfeld.

Christian, P. (1998). Soundex - can it be improved? *Computers in Genealogy*, 6(5):215–221.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string distance metrics for name-matching tasks. *IIWeb*, 3:73–78.

Cohen, W. W. (2000). Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems*, 18(3):288–321.

Gatschet, A. S. (1867). *Ortsetymologische Forschungen als Beiträge zu einer Toponomastik der Schweiz. 1*, volume 10, 325 S. Haller, Bern.

Jordan, P. (2012). Place names as ingredients of space-related identity. *Oslo Studies in Language*, 4(2):117–131.

Kondrak, G. (2005). N-Gram Similarity and distance. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Consens, M., & Navarro, G. (éd.), *String Processing and Information Retrieval*, volume 3772, pages 115–126. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kristol (2023). *Histoire linguistique de la Suisse romande volume 1 : De la préhistoire au Moyen Age*. Editions Alphil-Presses universitaires suisses, Neuchâtel.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Loup, R. & Bavaud, F. (2025). Spatial autocorrelation of political opinions: A kernel approach. *Journal of Geographical Systems*, pages 1–27.

Müller, W. (2002). Siedlungsgeschichte und Ortsnamen in der Suisse romande. In *Ortsnamen Und Siedlungsgeschichte*, pages 83–94. Regesta-Imperii, Heidelberg.

Odell, M. K. & Strong, E. P. (1947). *Records management and filing operations*, volume 9. McGraw-Hill Book Co., New York, NY.

Office Fédéral de la Statistique (2024). Application des communes suisses. https://www.agvchapp.bfs.admin.ch/fr.

Office Fédéral de la Topographie (2024). Répertoire officiel des localités. https://www.swisstopo.admin.ch/fr/repertoire-officiel-des-localites.

Vassere, S. (1996). 222. Morphologie et formation des microtoponymes: Domaine roman. In *Halbband+Registerband*, chapitre 2. Halbband+Registerband, pages 1442–1447. De Gruyter Mouton, Berlin, Boston, 2ème édition.

Annexes

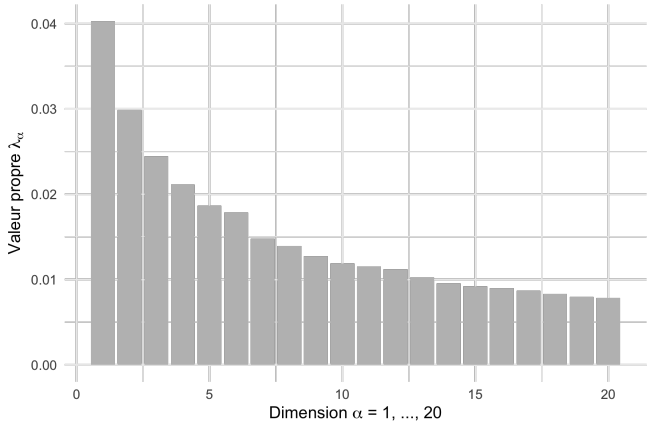


FIGURE 7 – *Scree graph* des 20 premières valeurs propres de $\mathbf{K}_{\text{nom}}$, pondéré par la taille des communes.

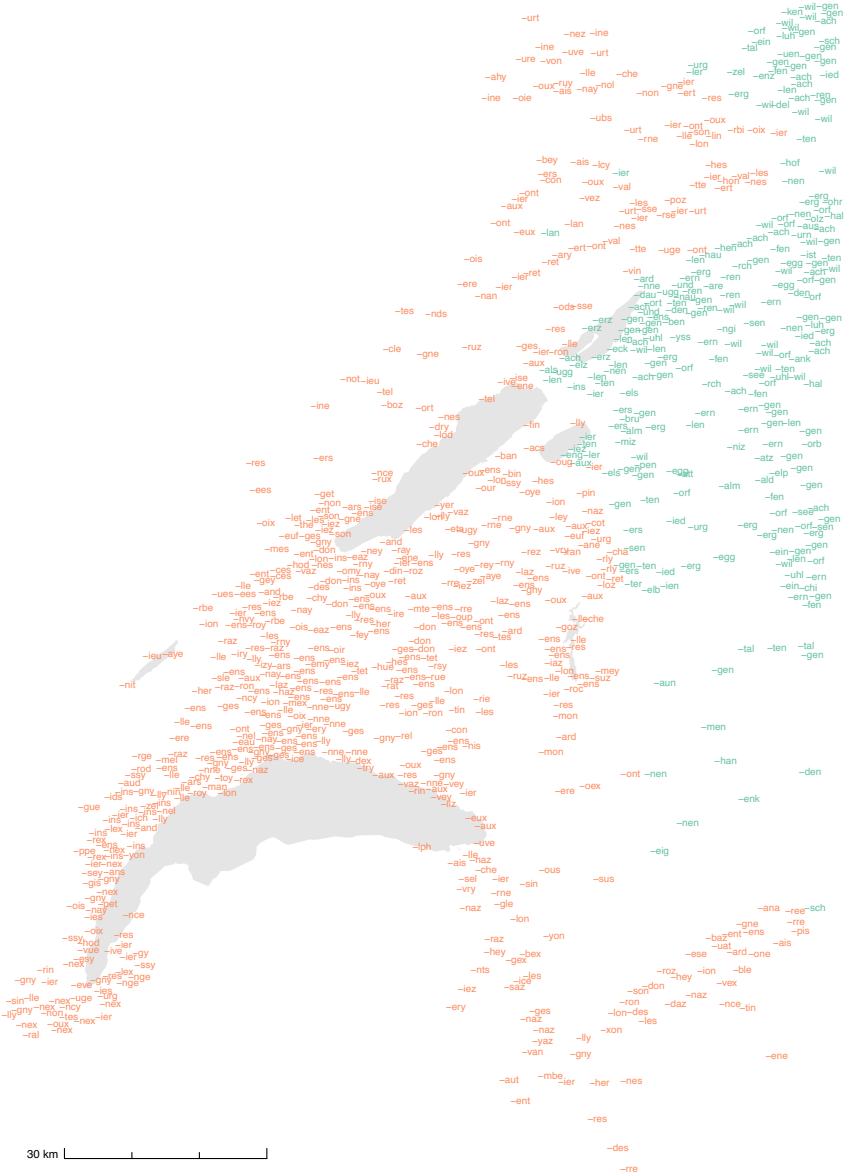


FIGURE 8 – Suffixes de la Romandie.

Cahiers du CLSL, n° 69, 2025, pp. 167-178

La vilaine *TINA* et le capitalisme autoritaire

Cédric Margot
Université de Lausanne
cedric.margot@unil.ch

Résumé

Cet article revisite l'argument principal de ma thèse : l'existence d'un *présupposé* de mesure qui – dans le discours de l'information – invalide de nombreuses affirmations statistiques. Le concept développé dans mon travail d'alors prend ici fonction rhétorique : je ne pose plus – dans une démarche inductive – la question de l'existence de ce présupposé ; je m'appuie, au contraire, sur la démonstration faite dans ma thèse et déduis qu'il permet de déconstruire des postures conservatrices, lorsque celles-ci reposent sur l'idée fausse que toute quantification peut se draper de l'objectivité de la mesure au sens strict. À ce titre, nous verrons que la célèbre formule de Margaret Thatcher « There is no alternative » exploite un présupposé de mesure et détache, par là même, les sciences économiques de leur épistémologie, celle des sciences humaines et sociales.

1 Introduction

Lorsque j'ai appris l'existence de ces mélanges en l'honneur de François Bavaud, j'ai tenu à y laisser une contribution sur un sujet connexe à ma thèse, qu'il connaît bien pour m'avoir fait l'honneur d'être un membre à la fois attentif et bienveillant du jury. La présente réflexion s'adresse à deux François : le scientifique passionné par son travail, et celui, plus engagé, que j'ai eu le plaisir – rare, mais toujours excellent – de côtoyer en d'autres occasions moins académiques. Dans ce texte, je revisite l'argument principal de ma thèse : l'existence d'un *présupposé de mesure* qui – dans le discours de l'information – invalide de nombreuses affirmations statistiques (Margot, 2023). Je m'attelle ici à

démontrer que la compréhension du concept de présupposé de mesure permet, à quiconque s'en empare, de déconstruire des postures conservatrices, lorsque celles-ci mobilisent tous azimuts des nombres pour camoufler leur caractère idéologique.

J'espère que François trouvera cette réflexion – née à la suite de discussions autour de mon colloque de thèse – intéressante, et que cette retraite – que je lui souhaite aussi longue et heureuse que possible – sera aussi l'occasion d'autres engagements, stimulés par une longue activité académique.

2 Rappel de la thèse

Constatant, d'une part, l'omniprésence des nombres et des statistiques dans le discours médiatique et, d'autre part, la diffusion d'une critique souvent peu informée – « on peut faire dire n'importe quoi aux chiffres » –, ma thèse interrogeait l'existence d'une confusion entre la mesure au sens strict et d'autres formes de quantification.

La capacité de raisonner objectivement sur les propriétés physiques de notre environnement par la mesure est une acquisition tardive dans l'histoire humaine. D'après Guénon (1945), « la mesure se rapporte principalement au domaine de la quantité continue » (p. 90), c'est-à-dire qu'elle présuppose la possibilité abstraite d'exprimer les grandeurs continues, ce qui ne fut rendu possible que par la découverte de la numération de position au V^e^ siècle de notre ère (Guedj, 1996; Ifrah, 1994) ; c'est en effet notre capacité à représenter les nombres réels qui permet d'envisager l'idée de précision.

D'autres avancées – essentielles dans la formation d'une définition moderne de la mesure – n'interviennent que bien plus tard. Ce n'est qu'au XIX^e^ siècle que des raisonnements mathématiques commencèrent à être systématiquement appliqués aux résultats des mesures dans l'ensemble des sciences de la matière. Selon l'épistémologue canadien Ian Hacking (1987), la période 1800-1850 marqua un tournant décisif avec le développement de méthodes permettant la mathématisation non plus seulement des sciences spatiales comme l'astronomie ou l'architecture, mais aussi des sciences dites « baconiennes », correspondant aujourd'hui à la physique, la biologie et la chimie (p. 48). À

partir de cette époque, les avancées dans la conception des instruments de mesure et le perfectionnement des techniques probabilistes favorisèrent une meilleure réplicabilité et une maîtrise accrue des marges d'erreur, établissant ainsi la mesure comme un outil d'objectivation fiable (Daston & Galison, 2021; Swijtink, 1987).

C'est à l'aune de ces évolutions qu'une définition de la mesure au sens étroit fut retenue dans le cadre de mon travail. Elle repose sur quatre piliers : (i) la possibilité théorique d'exprimer la continuité et l'infiniment précis grâce à la numération de position (par exemple les chiffres arabes), (ii) l'existence d'unités parfaitement définies et invariables (les unités de mesure), (iii) des instruments physiques de mesurage et (iv) des méthodes mathématiques permettant de maîtriser l'imprécision induite par l'activité de mesurage. Ces éléments réunis permettent à la mesure de produire des observations objectives ou « des observations sans observateur » pour citer l'historien et philosophe des sciences Zeno Switjink (1987). En d'autres termes, elle est une métrologie réaliste qui autorise l'expression de grandeurs continues (masse, distance, temps, ...) en nombres, sans faire intervenir la subjectivité des individus ; leur rôle se limite à la lecture du résultat d'un processus entièrement systématisé, dans lequel leur jugement n'intervient pas.

Face à l'efficacité avérée de la mesure dans les sciences dites « dures », la flexibilité des nombres fut progressivement adoptée dans le cadre des sciences humaines – notamment sous l'impulsion d'Adolphe Quetelet (1829; 1832a; 1832b; 1833). Cependant, les formes de quantification utilisées en sciences humaines et sociales, à la différence de la mesure au sens strict, reposent sur la définition de classes d'observation (quantités discrètes), lesquelles précèdent la traduction de l'expérience en chiffres (Desrosières, 1995). Il y a donc une étape de travail conventionnel intrinsèque à ces formes de passage du réel aux nombres ; celle-ci introduit immanquablement des éléments de subjectivité : les indices/indicateurs sont construits sur des choix et amènent, en quelque sorte, à intégrer l'observateur à ses observations[1].

1 À ce titre, le destinataire de cet hommage rappelle que les objets « idéaux » que sont les pièces de monnaie et les dés sont « éloignés des préoccupations des sciences humaines » (Bavaud, 2004, p. 5).

Ces différentes formes de passage du réel au chiffré produisent, les unes par rapport aux autres, des informations dont la valeur de vérité n'est pas la même du fait de leurs différences épistémologiques : la mesure est un transcodage fidèle des propriétés physiques du monde en nombres et offre des informations certaines (bien qu'imprécises), quand d'autres formes de quantification produisent des nombres qui n'ont toujours qu'un rapport probable à la vérité et reflètent un certain point de vue. Cela veut dire que les nombres issus de ces différents processus n'auraient pas vocation à être utilisés de la même manière dans la production d'information.

L'analyse d'un corpus de plus de 6 heures de téléjournaux, m'a permis de mettre en lumière que les objets quantifiés dans des champs des sciences humaines – notamment en économie – sont souvent traités, dans le discours de l'information, comme s'ils étaient obtenus par une métrologie réaliste, ce que j'ai nommé *présupposé de mesure* ; il apparaît lorsque des nombres obtenus par un processus de quantification incluant des étapes conventionnelles de définition sont traités comme s'ils étaient des entités objectives et neutres, à même de produire des informations certaines.

Un reportage (Radio Télévision Suisse, 2022) sur l'évolution du taux de chômage en Suisse au mois de janvier 2022 (exemple mobilisé dans ma thèse) illustre à merveille la réalisation de ce présupposé. Le calcul de ce taux à travers l'indicateur du Secrétariat d'État à l'Économie (SECO) offre une vision restrictive du chômage et est influencé par divers facteurs contextuels ; il est en effet basé essentiellement sur les inscriptions aux offices régionaux de placement (ORP). C'est notamment l'impact d'une loi sur l'obligation – dans certaines circonstances – d'annoncer les postes vacants en priorité auprès des personnes inscrites dans des ORP (LEI, RS 142.20) (Secrétariat d'Etat à l'Économie, 2022, p. 8) qui invalide cet indicateur : il est possible de montrer que cette loi a significativement modifié la dynamique des inscriptions aux ORP, notamment après la crise sanitaire. Cet état de fait autorise à effectuer une critique de l'interprétation des chiffres du SECO comme une mesure objective du marché du travail, et à relever un présupposé de mesure qui assimile ces statistiques à une réalité

tangible, alors qu'elles sont construites sur des bases conventionnelles et évolutives.

Malgré les éléments ci-dessus, le reportage de la RTS analysé dans la thèse contient des affirmations péremptoires qui laissent présupposer l'utilisation d'une métrologie réaliste typique des sciences exactes. S'y trouvent ainsi réunis les énoncés suivants « la situation s'améliore sur le marché du travail en Suisse », « le chômage est resté stable en janvier », « le marché du travail résiste très bien » et « le nombre de places vacantes [est] synonyme de dynamisme sur le marché du travail ». Ces énoncés se vident de leur sens et perdent toute valeur de vérité, dès lors qu'il est possible de montrer que l'évolution du taux de chômage constatée en janvier 2022 s'explique en grande partie par des effets mécaniques agissant plus sur les classes d'observation mobilisées par l'indicateur du SECO que sur l'état du marché du travail[2].

3 Réflexion épistémologique

Ma thèse doctorale – rédigée en section de linguistique – a été ancrée principalement dans le domaine de la *socio*linguistique : une discipline bien servie par son nom. En effet, elle appréhende les faits de langue en relation avec des faits sociaux ; c'est-à-dire qu'elle accorde une importance prépondérante au contexte de l'énoncé et au contexte social des locuteurs dans l'étude des faits linguistiques. C'est le cas au moins depuis les constats de l'un des « pères » de la discipline, Labov, résumés parfaitement dans ce passage de *Sociolinguistic Patterns* :

> [...] l'analyse du langage hors contexte, en tant que domaine autonome, subsistera sans aucun doute : comme toujours, il y

2 À cet égard, dans mon travail, je formulais l'hypothèse que « lorsque les effets de la pandémie se [seraient] totalement estompés, l'indicateur du SECO [repartirait], tout aussi mécaniquement à la hausse, le nombre de postes prioritaires pour les personnes inscrites [devant se réduire] à nouveau » (Margot, 2023, p. 346). Cette conjecture se trouve rétroactivement confirmée par les faits : la courbe du chômage au sens du SECO a continué sa chute induite par l'obligation d'annonce jusqu'à la fin de la crise sanitaire, puis s'est inversée à la hausse dans les derniers mois de 2022 et a constamment augmenté au cours des années 2023 et 2024, passant de 1.9% en septembre 2022 à 2.6% à la fin de 2024 (Secrétariat d'État à l'Économie, 2025). Cela pourrait être le fruit du hasard, mais j'aime à penser qu'une partie de mon analyse était pertinente.

> aura des linguistes pour consacrer tout leur temps à l'analyse de leurs intuitions sur la langue, et d'autres pour étudier les textes ou pour conduire des expériences de laboratoire [...] Mais désormais, la théorie linguistique ne pourra pas plus dédaigner le comportement social des sujets parlants que la chimie ne peut ignorer les propriétés observables des éléments. (Labov, 1972, pp. 350-351 ; traduit de l'anglais)

La sociolinguistique est ainsi une discipline particulière de la linguistique qui nomme son appartenance à l'épistémologie des sciences humaines et sociales. Parmi les sociolinguistes, il existe d'ailleurs un consensus : il ne peut pas y avoir de linguistique qui ne soit également de la sociolinguistique, puisqu'il n'y a pas de langue sans locuteurs présents ou passés. Cette thèse est donc écrite depuis une discipline qui a une forte conscience de son appartenance à un champ particulier du savoir : les sciences humaines.

D'autres disciplines, comme l'économie ou, dans une moindre mesure, la médecine, ne se présentent pas spontanément comme des sciences humaines. C'est-à-dire que ces champs décrivent des faits économiques ou médicaux en omettant parfois qu'ils sont également liés à des faits sociaux. L'économie – en particulier – a tout fait pour paraître une science dite « dure » : elle est allée jusqu'à se manufacturer un faux prix *Nobel* venant renforcer cette idée qu'elle était de ces sciences « plus nobles et plus rigoureuses » (Offer & Söderberg, 2016).

Je formule ici l'hypothèse que le présupposé de mesure découle d'un détachement de certaines sciences humaines d'avec leur épistémologie ; ces domaines ont oublié – ou ont fait oublier – qu'ils ne sont pas concernés par la mesure au sens strict.

4 *TINA* ! Le plus célèbre présupposé de mesure ?

Abordons à présent le plus célèbre (et sans doute le plus méconnu) des présupposés de mesure. Il porte l'acronyme *TINA* pour *There is no alternative*. Cette affirmation – dont raffolait l'ancienne Première Ministre britannique (1979-1990) Margaret Thatcher – est resservie à chaque attaque contre des acquis sociaux : il n'y aurait actuellement aucune alternative au rehaussement de l'âge de la retraite en France ; la

Suisse n'a guère fait mieux, en présentant la réforme de l'AVS comme une nécessité[3].

Un autre exemple marquant dans l'histoire récente est celui de la crise financière de 2008 : de nombreux gouvernements européens ont justifié des coupes budgétaires et des réformes économiques drastiques en invoquant l'absence d'alternative. En particulier, la Troïka (FMI, BCE, Commission européenne) a imposé des mesures d'austérité sévères à la Grèce, à l'Espagne et au Portugal sous prétexte qu'il n'existait « pas d'autre solution » pour garantir la stabilité économique et éviter la faillite de ces États (Europaforum, 2012; Kalara, 2024).

Plusieurs économistes ont critiqué les politiques d'austérité mises en œuvre après la crise financière de 2008, arguant qu'elles ont aggravé la récession et augmenté les niveaux de dette publique. Par exemple, l'association française des « Économistes atterrés » – fondée en 2011 – a dénoncé les politiques néolibérales et proposé des alternatives axées sur la croissance et la justice sociale (Askenazy et al., 2010, p. 28). L'économiste français Raveaud (2018) déplore également le « cauchemar grec », montrant qu'à la suite de la crise de 2008, l'Islande – se trouvant dans une situation similaire – a effectué des choix stratégiques différents, refusant le sacrifice de ses dépenses publiques (pp. 66-67). De même, l'économiste nobélisé Stiglitz (2002) a critiqué les politiques d'austérité, estimant qu'elles ont souvent aggravé les problèmes économiques des pays concernés. Ces critiques montrent que l'argument *TINA* relève plus d'un choix idéologique que d'une nécessité économique, et qu'il existe bien des alternatives aux politiques d'austérité.

L'affirmation – selon laquelle il n'y aurait pas d'alternative – ne suppose-t-elle pas que cette absence d'option peut être décrétée de façon péremptoire à l'appui d'outils objectifs ? Or, ces injonctions se font sur la base d'indicateurs qui ne sont pas des outils de mesure au sens strict. Comme relevé, ces objets reposent sur un important travail de définition et de délimitation de classes. C'est-à-dire qu'il y a, en réalité, une multitude d'alternatives, puisque les métrologies utilisées ont un fondement conventionnel et ainsi, par définition, discutable.

3 Le rehaussement de l'âge de la retraite pour les femmes, associé à une hausse de la TVA pour financer les retraites.

L'idée qu'il n'y aurait pas d'alternative dans la résolution des problèmes économiques et sociaux présuppose ainsi la possibilité d'écarter d'autres approches par la mesure, puisque seule une métrologie réaliste – se fondant sur des classes d'observation invariables et indiscutées – pourrait permettre de postuler objectivement l'absence d'option. Par extension, dans le cadre des sciences humaines, postuler l'absence d'alternative est toujours et immanquablement un présupposé de mesure, où la prétendue mesure joue un rôle sophistique et sert d'argument d'autorité.

5 La rhétorique de l'impasse fondée sur l'outil statistique : une marque d'un capitalisme autoritaire

Ce discours de l'impasse – seul un retour en arrière serait possible – repose systématiquement sur des indicateurs statistiques. Mais ces indicateurs ne sont pas mobilisés comme de simples outils d'aide à la décision, permettant à celui qui tient le gouvernail de mieux orienter sa trajectoire. Ils sont, au contraire, détournés en arguments d'autorité, servant à justifier ce qui est présenté comme nécessaire et incontestable. Ce glissement – qui transforme un instrument de « mesure » en un instrument de conviction – est, selon plusieurs auteurs, parmi lesquels Cardon (2015), Bourdieu (1998) et Desrosières (2008), une caractéristique structurelle du néolibéralisme. Il transcende les cadres institutionnels et s'observe aussi bien dans un système à forte démocratie directe, comme la Suisse, que dans un système où le vote sur le fond est relativement rare, comme en France.

Loin d'être une simple dérive médiatique, ce phénomène participe d'une mécanique plus large de manufacture du consentement. Comme le souligne le philosophe français Jacques Rancière (2022), les médias ne sont pas tant là pour tromper les citoyens d'une démocratie que pour assurer la pédagogie d'un ordre dominant. Cet ordre, qui exerce une influence considérable sur les milieux politiques, détient aujourd'hui une part significative des médias (Le Monde Diplomatique & Acrimed, 2024) et des sources d'information, ce qui lui permet d'orienter les récits et d'imposer ses cadres d'analyse.

C'est dans ce contexte que le chiffre, posé comme un invariant, ac-

quiert une fonction de légitimation et de preuve. Puisque les journalistes ne participent ni à l'élaboration des statistiques ni, souvent, à leur interprétation critique, la confusion entre la mesure – une métrologie réaliste – et d'autres formes de quantification produit une illusion rhétorique : celle d'une absence totale d'alternative aux politiques mises en place. Il est à déplorer que, dans les médias, les chiffres ne soient presque jamais questionnés en tant que constructions sociales. Ils sont le plus généralement envisagés comme des faits, et non comme le résultat de choix méthodologiques et politiques. Cette dépolitisation de la quantification est au cœur de ma thèse : elle conduit à la production d'énoncés dépourvus de valeur de vérité, puisqu'ils reposent sur le présupposé erroné d'un recours systématique à une métrologie réaliste.

Cette mécanique de légitimation s'inscrit dans une dynamique plus vaste, que Foucault (2004a; 2004b) qualifiait de *gouvernementalité*. Dès les années 1980, il identifiait dans le néolibéralisme une nouvelle forme de pouvoir, fondée sur la gestion des conduites par des mécanismes d'incitation et d'autorégulation des individus. Mais à mesure que ce modèle s'impose, il nous entraîne dans une impasse intellectuelle conservatrice, au point que *les* alternatives sont réduites à une absence d'« alternative » – au singulier – dans la célèbre formule. *TINA* n'est-elle pas constitutive d'une attaque réactionnaire contre les libertés ? Cette prétendue absence d'alternative nous invite à mettre en question le caractère *libéral* de ce *néolibéralisme* : dans cette rhétorique, toutes les autres options idéologiques sont rendues inenvisageables.

Cette interrogation est aujourd'hui largement partagée. Comme l'a récemment souligné Bégaudeau (2023), le terme « néolibéralisme » est devenu une catégorie floue, omniprésente mais difficilement définissable, cela en dépit du fait que « certains s'arment de rigueur pour donner de la consistance au mot en réduisant son champ » (p. 129). Il trouve un « défaut de fabrication » à l'appellation : « dans néolibéralisme on entend libéralisme dans quoi on entend liberté [...] » Pour lui, « néolibéralisme est une néo-manière de ne pas dire capitalisme » ; il propose ainsi une ré-appellation permettant de « synthétiser des faits matériels » (p. 132) et de s'éloigner de cette notion vague et fourre-tout : *néocapitalisme*.

La remise en question du terme *néolibéralisme* ne se limite évidemment pas au cadre littéraire et se retrouve dans les milieux académiques. La philosophe française Stiegler (2019) parle désormais de *nouveau libéralisme autoritaire*. La professeure étasunienne en sciences politiques Brown (2015) insiste sur le fait qu'il s'agit d'une reconfiguration de la manière dont les individus se perçoivent, interagissent et sont gouvernés ; elle préfère parler de *gouvernementalité néolibérale* pour insister sur la nature performative et insidieuse de cette forme de pouvoir. De son côté, Chamayou (2018) – chercheur au CNRS – décrit comment les élites économiques ont développé un *libéralisme autoritaire*, conçu comme une stratégie visant à neutraliser les contestations démocratiques et à renforcer le pouvoir des grandes entreprises. Par ailleurs, plusieurs économistes – comme Raphaël Rossello, Gaël Giraud et Gilles Raveaud (Thinkerview, 2022) – soulignent que ceux qui se revendiquent du libéralisme aujourd'hui semblent être devenus alibéraux. Ces auteurs convergent vers la même observation : la réduction des libertés individuelles apparaîtrait comme une caractéristique de ce modèle qui n'est ainsi *libéral* que sous certains aspects.

Dans ce contexte, le concept de *présupposé de mesure* me semble particulièrement pertinent. En déconstruisant une rhétorique de l'impasse – classiquement fondée sur des arguments statistiques – et en mettant en évidence la confusion entre quantification et mesure, il permet d'expliquer, sur une base linguistique et scientifique, comment certains discours – notamment ceux visant à nous faire accepter une absence d'alternative – se situent hors du champ de la vérité. Puissions-nous continuer à travailler à une alternative, comme notre François a eu à cœur de le faire en enseignant, année après année, les dangers d'une prétention à l'évidence.

Références

Askenazy, P., Coutrot, T., Orléan, A., & Sterdyniak, H. (2010). *Manifeste d'économistes atterrés*. Les Liens qui Libèrent, Paris.

Bavaud, F. (2004). *Modèles et données : une introduction à la statistique uni-, bi- et trivariée*. L'Harmattan, Paris.

Bégaudeau, F. (2023). *Boniments*. Multitudes, Amsterdam.

Bourdieu, P. (1998). L'essence du néolibéralisme. *Le Monde diplomatique*. https://www.monde-diplomatique.fr/1998/03/BOURDIEU/3609.

Brown, W. (2015). *Undoing the demos: Neoliberalism's stealth revolution*. Near Futures. Zone Books, New York.

Cardon, D. (2015). *A quoi rêvent les algorithmes: nos vies à l'heure des « big data »*. Seuil, Paris.

Chamayou, G. (2018). *La société ingouvernable. Une généalogie du libéralisme autoritaire*. La Fabrique Éditions, Le Kremlin Bicêtre.

Daston, L. & Galison, P. L. (2021). *Objectivity*. Princeton University Press.

Desrosières, A. (1995). Classer et mesurer : les deux faces de l'argument statistique. *Réseaux. Communication - Technologie - Société*, 13(71):11–29.

Desrosières, A. (2008). *Pour une sociologie historique de la quantification : l'argument statistique I*. Sciences Sociales. Presses des Mines, Paris.

Europaforum (2012). Les représentants de la troïka sont venus expliquer aux eurodéputés qu'il n'y a pas d'alternative à la « potion amère » prescrite à la Grèce, tout en appelant les autorités grecques à faire preuve de « courage politique » pour mener les réformes. https://europaforum.public.lu/fr/actualites/2012/03/pe-troika-grece/index.html. Consulté le 1er janvier 2025.

Foucault, M. (2004a). *Naissance de la biopolitique : cours au Collège de France (1978-1979)*. Cours au Collège de France / Michel Foucault [8]. Gallimard, Paris.

Foucault, M. (2004b). *Sécurité, territoire, population : cours au Collège de France (1977-1978)*. Cours au Collège de France / Michel Foucault [7]. Gallimard, Paris.

Guedj, D. (1996). *L'empire des nombres*. Gallimard, Paris.

Guénon, R. (1945). *Le règne de la quantité et les signes des temps*. Gallimard, Paris.

Hacking, I. (1987). Was there a probabilistic revolution 1800-1930? In *The probabilistic revolution: ideas in history*, volume 1, pages 45–55. The MIT Press, Cambridge, MA, US.

Ifrah, G. (1994). *Histoire universelle des chiffres: l'intelligence des hommes racontée par les nombres et le calcul*. Robert Laffont, Paris.

Kalara, M. (2024). Crise économique et mauvaises pratiques de législation : la Grèce comme laboratoire d'expérimentation constitutionnelle ? *La Revue des Droits de l'Homme*, 25.

Labov, W. (1972). *Sociolinguistic patterns*. Conduct and Communication, 4. University of Pennsylvania Press, Philadelphia.

Le Monde Diplomatique & Acrimed (2024). Médias français, qui possède quoi ? https://www.monde-diplomatique.fr/cartes/PPA. Consulté le 31 jan. 2025.

Margot, C. (2023). *L'ombre de la quantification : le présupposé de mesure dans le discours de l'information*. Thèse de doctorat, Université de Lausanne.

Offer, A. & Söderberg, G. (2016). *The Nobel factor*. Princeton University Press.

Quetelet, A. (1829). *Recherches statistiques sur le Royaume des Pays-Bas*. Hayez, Bruxelles.

Quetelet, A. (1832a). Recherches sur la loi de la croissance de l'homme. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 7.

Quetelet, A. (1832b). Recherches sur le poids de l'homme aux différens âges. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 7.

Quetelet, A. (1833). *Recherches sur le penchant au crime aux différens âges*. Hayez, Bruxelles.

Radio Télévision Suisse (2022). Téléjournal de 12h45. https://www.rts.ch/play/tv/12h45/video/12h45?urn=urn:rts:video:12849930. 8 février 2022.

Rancière, J. (2022). Les désordres du monde. https://www.youtube.com/watch?v=zrBekCSf_-8. [Entretien].

Raveaud, G. (2018). *Economie : on n'a pas tout essayé !* Seuil, Paris.

Secrétariat d'Etat à l'Économie (2022). Monitorage relatif à l'exécution de l'obligation d'annoncer les postes vacants. https://www.arbeit.swiss/secoalv/fr/home/menue/unternehmen/stellenmeldepflicht.html.

Secrétariat d'État à l'Économie (2025). Chiffres du chômage. https://www.seco.admin.ch/seco/fr/home/Arbeit/Arbeitslosenversicherung/arbeitslosenzahlen.html. Consulté le 2 avr. 2025.

Stiegler, B. (2019). *« Il faut s'adapter » : sur un nouvel impératif politique*. Gallimard, Paris.

Stiglitz, J. E. (2002). *La grande désillusion*. Fayard, Paris.

Swijtink, Z. G. (1987). The Objectification of Observation. In *The Probabilistic Revolution: Ideas in History*, volume 1, pages 261–285. The MIT Press, Cambridge, MA, US.

Thinkerview (2022). Crise financière: la descente aux enfers ? Gaël Giraud, Raphaël Rossello & Gilles Raveaud. https://www.youtube.com/watch?v=n7oj2m8B0iM. [Entretien].

Cahiers du CLSL, n° 69, 2025, pp. 179-188

πάντα ῥεῖ : changements sémantiques dans la terminologie mathématique

Robin Meyer
Université de Lausanne
robin.meyer@unil.ch

Résumé

Tout le monde est témoin du fait que les langues changent au fil du temps quant au lexique dont elles se servent (p. ex. *ouèch*, *kiffer*, *mdr* en français contemporain) ou en ce qui concerne les structures qu'elles permettent (p. ex. l'emploi du subjonctif après *après que* en analogie avec *avant que*). Pour la plupart, ces changements sont motivés (et se retrouvent) dans la langue quotidienne et informelle où l'on accorde moins d'attention au soin qu'à l'écrit.

Cependant, même des termes de la science, justement considérés immuables de nos jours, ont subi de tels changements avant d'être stabilisés dans leur forme actuelle. Tout au début, les mots à l'origine des termes *mathématique* et *statistique*, par exemple, n'avaient rien à faire avec ce qu'ils dénotent aujourd'hui : en premier lieu, le mot grec μαθηματικός signifiait « concernant la matière qui est apprise » alors que l'italien *statistica* se référait d'abord à la « science de (la gestion de) l'état ».

Ce sont ces changements dans la terminologie scientifique, leur origine, déroulement et motivation dont traitera cette petite contribution en l'honneur de notre cher collègue.

1 Introduction

Depuis au moins la deuxième moitié du XIX^{e} siècle, les linguistes et philologues commencèrent à se servir des méthodes mathématiques pour enrichir et étoffer leurs recherches ou corroborer leurs hypothèses. Au début, il s'agissait surtout de simples comptages de mots ou de

constructions ;[1] de nos jours, cependant, avec l'avènement des méthodes statistiques de plus en plus complexes, aidées par les capacités des ordinateurs, plusieurs domaines de la linguistique, comme l'évaluation des corpus ou la linguistique forensique, dépendent de la mathématique alors que d'autres, tels que la sociolinguistique ou la linguistique historique, l'utilisent parmi d'autres outils. L'exploration scientifique moderne du langage et des langues n'est donc guère concevable sans accès à la mathématique.

Or, cette dépendance est largement inégale en ce sens que la mathématique ne se servit que rarement des instruments ou méthodes linguistiques, à l'exception peut-être de la création de nouveaux termes. Toutefois, une vraie appréciation de la mathématique exige aussi la connaissance de son histoire, non seulement en ce qui concerne le développement de la pensée scientifique, mais aussi de la terminologie utilisée.

En l'honneur de notre cher collègue, cette petite étude revient donc sur l'origine et l'histoire de deux termes fondamentaux pour chaque mathématicien·ne appliqué·e, ceux de la *mathématique* elle-même et de la *statistique*, et propose une esquisse de leur développement en français, en anglais et dans l'Antiquité, dans la mesure du possible. Il s'avère que les deux termes ont subi, à différents degrés, des changements sémantiques bien attestés.

2 *Mathématique*

Afin d'explorer dans quelle mesure le sens du mot *mathématique* a changé au fil du temps, il faut en premier lieu adopter une définition moderne. Malheureusement, la littérature scientifique dédiée au sujet ne s'aventure pas à définir des termes si élémentaires même si elle tente d'en discuter l'histoire.[2] Il ne reste donc que la définition de

1 La question du phénomène de l'*attractio relativi*, par exemple, est discuté sur cette base par Förster (1868) ; cependant, le même phénomène est étudié de nos jours avec des méthodes (au moins un peu) plus avancées, cf. Probert (2015); Meyer (2018).

2 Cf. p. ex. Kendall (1960) ; Perisho (1965, p. 64) ; Lo Bello (2013, pp. 199-207). Ce dernier, malgré sa prétention d'exhaustivité (« comprehensive dictionary »), semble omettre nombre de termes importants et donner des explications parfois assez biaisées (cf. Hollings 2024).

dictionnaires : en français, le *TLFi* distingue plusieurs usages du mot, comme nom et adjectif, y compris des abstractions comme « qui est purement abstrait » ; tout au fond cependant, se trouve l'idée que les mathématiques (au pluriel) sont

> [l'e]nsemble des disciplines qui procèdent selon la méthode déductive et qui étudient les propriétés des êtres abstraits comme les nombres, les figures géométriques ainsi que les relations qui existent entre eux.[3]

L'*OED* propose une définition qui n'est pas trop éloignée, mais montre néanmoins des différences remarquables :

> Originally : (a collective term for) geometry, arithmetic, and certain physical sciences involving geometrical reasoning, such as astronomy and optics ; spec. the disciplines of the quadrivium collectively.
> In later use : the science of space, number, quantity, and arrangement, whose methods involve logical reasoning and usually the use of symbolic notation, and which includes geometry, arithmetic, algebra, and analysis ; mathematical operations or calculations.[4]

Les deux approches soulignent l'importance d'une certaine méthode de raisonnement (dite déductive ou logique) ainsi que la compositionnalité de la discipline, qui en réunit de nombreuses autres. Cette perspective date de l'Antiquité (voir ci-dessous) et fait également partie des premières définitions trouvées dans la littérature française, comme par exemple dans la discussion de « toutes les sciences » (théologie, physique et mathématique) dans *Li Livres dou Tresor* (1265) de Brunetto Latini (ca. 1220–1294), érudit florentin et ancien tuteur de Dante Alighieri, qui résume de façon encyclopédique l'état des connaissances de ce temps-là :

> La tierce [escience] est matematique, por cui nos savons les natures des choses qui n'ont point de cors & sont entor les

3 *TLFi* (2024, s.v. mathématique).
4 *OED* (2024, s.v. mathematics (n. 2)).

> corporaus choses. [...] ce sont .iiii. escienses el cors de matematique ; & sont apelees por droit nom arismetique, musique, jemetrie, astronomie.[5]

En anglais, la première définition du mot (et presque la plus ancienne attestation aussi) date d'environ 1393 et se trouve dans le poème *Confessio amantis* de John Gower (ca. 1330–1408), poète anglais et contemporain de Geoffrey Chaucer :

> The thridde point of Theorique, Which cleped is Mathematique, Devided is in sondri wise[...] The ferste of whiche is Arsmetique, And the secounde is seid Musique, The thridde is ek Geometrie, Also the ferthe Astronomie.[6]

Le développement définitoire le plus frappant dans l'histoire du français et de l'anglais (et sans doute dans la vaste majorité d'autres langues modernes) consiste en la scission de certaines sciences auparavant considérées comme appartenant à la mathématique, comme la musique et l'astronomie, bien que ces dernières continuent évidemment à utiliser des méthodes mathématiques et il y ait d'autres recoupements entre ces (sous-)domaines. Ce type de changement n'est guère surprenant, sachant que les sciences dans le sens moderne faisaient partie du champ de la philosophie (dans le sens pré-moderne) pour la plupart de l'histoire humaine.

Ce type de changement sémantique, dit restriction, est bien attesté dans l'histoire des langues du monde. L'anglais *starve* « mourir de faim », par exemple, vient du vieil anglais *steorfan*, qui est cognat avec l'allemand *sterben* « mourir » (du proto-germanique **sterbaną*) ; l'anglais a donc restreint le sens alors que l'allemand a retenu la version plus générale. Pareillement, en français moderne le mot *chômage* se réfère surtout à la condition sans emploi, une restriction du sens simple du verbe *chômer* « cesser une activité ; ne pas travailler », dérivé à son tour du latin tardif *caumāre* « faire une pause pendant la chaleur de midi » (du grec ancien καῦμα « chaleur »).

Une évolution semblable avait eu lieu dans l'Antiquité déjà. L'origine direct du terme dans les langues modernes se trouve en latin, où

5 Baldwin & Barrette (2003, pp. 3-4).
6 VII.145-152 (Macaulay, 1901, III. pp. 237).

existaient déjà l'adjectif *mathēmaticus* ainsi que la désignation de la science elle-même, *mathēmatica*. Vitruve parle des *mathēmatica nota* « signes mathématiques » ;[7] Aulu-Gelle lie l'usage du mot latin à son origine en grec :

> quoniam geometriam, gnomonicam, musicam ceterasque item disciplinas altiores μαθήματα veteres Graeci appellabant
> « puisque les grecs anciens appelaient la géométrie, la gnomonique,[8] la musique et d'autres études élevées μαθήματα ou "sciences" ».[9]

Alors que les Romains aussi avaient emprunté le terme, la définition d'Aulu-Gelle montre que la notion médiévale revient à celle de l'Antiquité, où la mathématique incluait d'autres disciplines qui de nos jours sont indépendantes.

La vraie origine, comme déjà indiqué ci-dessus, de la mathématique (au moins dans le sens étymologique) réside donc dans la langue grecque. L'adjectif μαθηματικός *mat^{h}ēmatikós* fait référence à la science elle-même, aux individus concernés par cette science, et à différentes sous-disciplines telles que l'astronomie, comme déjà mentionné. Pourtant, l'une des premières attestations du mot, qui se trouve dans le dialogue *Timaeus* de Platon, remet en question si l'équivalence entre μαθηματικός et *(qui étudie la) mathématique* tient déjà à ce temps-là :

> τὸν δὴ μαθηματικὸν ἤ τινα ἄλλην σφόδρα μελέτην διανοίᾳ κατεργαζόμενον καὶ τὴν τοῦ σώματος ἀποδοτέον κίνησιν, γυμναστικῇ προσομιλοῦντα
> « Ainsi l'étudiant de la mathématique ou quiconque exerce beaucoup son intellect doit également entraîner son corps, en faisant de la gymnastique. »[10]

7 Vitr. *De arch.* 1. 1. ; les citations des œuvres de l'Antiquité suivent les conventions notées dans l'*Oxford Classical Dictionary*.
8 La gnomonique est la science qui concerne la conception, le calcul et la manufacture des cadrans solaires.
9 Gell. *NA* 1.9.6.
10 Pl. *Ti.* 88c.

Il semble plutôt que la signification dans ce contexte devrait être plus large, plus générale, peut-être dans le même sens que l'adjectif cognat μαθητικός *mat*^{h}*ētikós* « qui aime l'érudition/la connaissance ». Les deux formes sont en tout cas dérivées de la même racine °*mat*^{h}- « apprendre, connaître » attestée aussi dans d'autres langues de la famille indo-européenne.[11]

Il s'avère donc, en passant en revue ces adjectifs ainsi que le nom de base associé (μάθημα *mát*^{h}*ēma* « ce que l'on apprend ; érudition, connaissance ; science »), que tout au début, la mathématique était, dans un sens, *la* science par excellence. Cependant, comme illustré, ce sens large fut limité à différentes reprises.

3 *Statistique*

L'histoire du terme *statistique*, en revanche, montre des changements opposés à ceux mentionnés ci-dessus : dans ce cas, le sens du mot n'a pas été précisé, mais plutôt étendu.

De nouveau, comme point de départ, certaines définitions servent moins que d'autres. Lo Bello propose la suivante :

> This is a very low word of the eighteenth century, the offspring of a succession of mistakes. The formula for it is Latin *status* + Greek -ιστής + Greek -ικός + English *s*. The name comes from the fact that the collection of data was originally an activity of the state.

Cette définition est suivie par un extrait du Nouveau Testament (Lc. 2 : 1), qui parle du recensement conduit en 6 apr. J.-C. sous la légation de Publius Sulpicius Quirinius (ca. 51 av. J.-C. – 21 apr. J.-C.) en tant que première « statistique » communément connue.

L'étymologie proposée, malgré le caractère inutilement évaluatif de cette remarque, est fondamentalement correcte,[12] mais cache une histoire plus complexe. Premièrement, il faut souligner que de telles combinaisons de latin et grec ancien (dites *composita mixta*), dont se

11 Cp. lat. *mēns* « esprit, intellect » ; skt. *matá* « pensée, croyance ».

12 *Pace* Missiakoulis (2024) qui propose un lien étroit avec grec ancien στάσις *stásis* « position, positionnement » et στατός *statós* « mis, positionné » sans pouvoir motiver le changement sémantique.

moque Lo Bello, sont bien établies dans l'histoire de plusieurs langues d'Europe, p. ex. *automobile* (grec αὐτός *autós* « soi-même » + lat. *mobilis* « mobile, agile ») ou *néonatal* (grec νέος *néos* « nouveau, neuf » + lat. *natalis* « relatif à la naissance »), et ne méritent pas d'attention particulière.

Aucune source ancienne ne se sert d'un mot d'une telle composition, raison pour laquelle il faut partir du principe qu'il s'agit d'une création moderne. La première attestation se trouve au début du XVII^e^ siècle dans l'œuvre de l'écrivain et érudit italien Girolamo Ghilini (1589–1668), qui en 1633 parlait de *statistici affari* et *scienza statistica*.[13] Or, en ce temps-là, cet adjectif signifiait plutôt « relatif à l'état », comme le contexte le trahit. Cet usage continue également ailleurs, comme dans la première attestation allemande en 1749 dans le traité *Abriß der neuen Staatswissenschaft der vornehmen Europäischen Reiche und Republiken* du philosophe et économe Gottfried Achenwall (1719–1772). Il admet que le nouveau terme était encore assez vague : « Der Begriff der sogenannten Statistic, das ist, der Staatswissenschaft einzelner Reiche wird sehr verschiedentlich angegeben ».[14] C'est par la médiation de l'allemand que le mot s'est inséré en français, puis en anglais.

Quant au sens, cette signification plutôt politique ne passe pas directement à la version actuelle sans détournement. Alors que le terme courant est défini dans l'*OED* comme

> [t]he systematic collection and arrangement of numerical facts or data of any kind; (also) the branch of science or mathematics concerned with the analysis and interpretation of numerical data and appropriate ways of gathering such data,[15]

il y a une phase de transition sémantique attestée dans les deux langues. Tant le français que l'anglais connaissent un sens intermédiaire entre « (science) relative à l'état » et « analyse et interprétation de données, quantifiables », un sens qui traite de l'« [é]tude méthodique des faits économiques et sociaux par des classements, des inventaires chiffrés, des

13 Cf. Missiakoulis (2024) et Ostasiewicz (2014) pour le débat du lieu d'attestation.
14 Achenwall (1749, §1).
15 *OED* (2024, s.v. statistics (n. 1.b.)).

recensements, etc. ».[16] Tel est l'usage dans l'*Essai sur les fondements de nos connaissances et sur les caractères de la critique philosophique* (1851) d'Antoine Augustin Cournot (1801–1877), mathématicien et philosophe français, qui utilise le mot fréquemment dans ce sens, p. ex. ici :

> La statistique apprendra bien que la population s'est accrue, que le prix des denrées a baissé ou haussé[17]

Une telle signification est attesté en anglais depuis 1800 ; l'exemple le plus vieux qui mentionne clairement la connexion entre chiffres et gestion de l'état se trouve chez Harriet Martineau (1802–1876), sociologue britannique, qui en traite dans sa *Society in America* de 1837 :

> There is great virtue in figures, dull as they are to all but the few who love statistics for the sake of what they indicate.[18]

La définition moderne se développe au cours du XIX^e^ siècle sans qu'il soit (au niveau des modestes connaissances de l'auteur de la présente contribution) possible d'établir sans équivoque quelle était la première occurrence, les frontières entre les deux derniers sens étant fluides.

Les changements sémantiques impliqués dans l'évolution de ce terme, de « science de (la gestion de) l'état » à « branche de la mathématique qui traite de l'analyse de données » via « étude des données économiques », peut donc être décrit simplement (et *grosso modo*) comme une extension, l'opposé de la restriction mentionnée pour le terme *mathématique*. Ici, le sens du mot est devenu de moins en moins spécifique pour, à la fin, englober l'analyse de tout type de données (de manière mathématique). De nouveau, de tels changements sont bien attestés, comme par exemple dans le mot français *doyen*, dérivé du latin *decānus* « dizenier », parce qu'un doyen n'est plus juste une désignation militaire et peut avoir plus de dix personnes sous sa direction ; pareillement, le verbe français *arriver* se base sur le latin tardif *arrīpāre* « descendre à terre (depuis un bateau) » (< *ad* « à » + *rīpa* « rive ») et montre aussi clairement ce type de changement sémantique.

16 *TLFi* (2024, s.v. statistique) ; cp. *OED* (2024, s.v. statistics (n., 3.a.)).
17 Cournot (1851, II. p. 251)
18 Martineau (1837, II. p. 292).

4 Remarques finales

Il va sans dire que cette contribution n'a pas pu être exhaustive et que, du point de vue à la fois de la sémantique historique et de la lexicographie, de nombreuses bonnes questions mériteraient une considération ultérieure, comme par exemple la régularité (ou non) des abréviations des deux termes traités ici en français et en anglais (*maths*, *stats*) et la question de la pluralisation en anglais et français (*mathematics* et *mathématiques*, *statistics*).

Malgré ce défaut, reste l'espoir que ce « tourbillon » d'une histoire sémantique des termes mathématiques plaise à notre cher collègue et que nous ayons le temps d'en parler davantage.

Références

Achenwall, G. (1749). *Abriß der neuesten Staatswissenschaft der vornehmsten Europäischen Reiche und Republicken*. Johann Wilhelm Schmidt, Göttingen, 1ère édition.

Baldwin, S. & Barrette, P. (2003). *Brunetto Latini, li livres dou Tresor. Edition and study*. Arizona Center for Medieval and Renaissance Studies, Tempe, AZ.

Cournot, A. A. (1851). *Essai sur les fondements de nos connaissances et sur les caractères de la critique philosophique*. Hachette, Paris.

Förster, R. (1868). *De attractione enuntiarum relativarum qualis quum in aliis tum in graeca lingua potissimumque apud graecos poetas fuerit*. Mitscher und Röstell, Berlin.

Hollings, C. (2024). Review of Lo Bello (2013). https://ima.org.uk/369/origins-of-mathematical-words-a-comprehensive-dictionary-of-latin-greek-and-arabic-roots/.

Kendall, M. G. (1960). Studies in the history of probability and statistics. where shall the history of statistics begin? *Biometrika*, 47(3/4):447–449.

Lo Bello, A. (2013). *Origins of mathematical words. A comprehensive dictionary of Latin, Greek, and Arabic Roots*. Johns Hopkins University Press, Baltimore.

Macaulay, G. C. (1899–1901). *The complete works of John Gower, edited from the manuscripts with introductions, notes, and glossaries*. Clarendon Press, Oxford.

Martineau, H. (1837). *Society in America*. Saunders and Otley, New York, NY.

Meyer, R. (2018). Syntactical peculiarities of relative clauses in the Armenian New Testament. *Revue des Études Arméniennes*, 38:35–83.

Missiakoulis, S. (2024). Some linguistic aspects of the term "statistics". *Encyclopedia*, 4(3):1286–1291.

Ostasiewicz, W. (2014). The emergence of statistical science. *Śląski Przegląd Statystyczny*, 12(18).

Perisho, M. W. (1965). The etymology of mathematical terms. *Pi Mu Epsilon Journal*, 4(2):62–66.

Probert, P. (2015). *Early Greek relative clauses*. Oxford University Press, Oxford.

OED (2024). Oxford english dictionary. https://www.oed.com/.

TLFi (2024). Trésor de la langue française informatisé. http://atilf.atilf.fr/m.

Cahiers du CLSL, n° 69, 2025, pp. 189-206

Le doigt et les étoiles

Isaac Pante
Université de Lausanne
isaac.pante@unil.ch

Résumé

Cet article interroge les procès de légitimité auxquels doit régulièrement faire face l'informatique en Faculté des lettres. En prenant le cas particulier de la Section des sciences du langage et de l'information de l'Université de Lausanne (UNIL), il fait l'histoire d'un certain nombre de stratégies d'intégration, présente leurs limites, et invite la section à se doter d'un corpus d'œuvres qui lui soit propre, à l'instar des disciplines canoniques en Faculté des lettres. Le jeu (vidéo) est présenté comme un candidat de choix : objet hybride, son analyse suppose de prendre en considération ses dimensions tant statistiques que narratives. L'article illustre le bénéfice d'un double regard computationnel et littéraire appliqué au jeu comme forme culturelle et liste les enjeux des graphes et des réseaux pour l'enseignement et la recherche en fiction interactive, qu'il s'agisse d'analyser les mécaniques des *Livres dont vous êtes le héros*, de créer des textes non linéaires ou de développer des compétences en game design.

1 Contextualisation

1.1 Du côté du diable

L'année dernière, l'informatique pour les sciences humaines fêtait ses trente ans d'existence à l'UNIL. Son histoire houleuse mériterait un roman. Après tout (comme n'ont pas manqué de le demander collègues et étudiant·e·s au fil des ans), que diable fait l'informatique en Faculté des lettres ?

À l'heure des humanités numériques et de l'intelligence artificielle, cette diablerie – le fameux « pacte » de Moretti (Moretti, 2000, 57) –

émeut bien moins qu'autrefois. Sans s'adonner à une futurologie périlleuse, on peut même penser que l'informatique et les méthodes computationnelles finiront par se diffuser dans toutes les disciplines, à l'instar de la linguistique à l'ère triomphante du structuralisme. D'aucun·e·s espèrent d'ailleurs que l'incitation à développer les compétences numériques dans toutes les formations (voulue par la Confédération et la plupart des instances de gouvernance fidèles au *New Public Management*) signera la fin des interminables procès en légitimité de l'informatique en-dehors des cursus de pure ingénierie.

Ce serait pourtant bien mal connaître l'histoire des sciences. Contre toute attente (du moins pour qui ignore la psychologie sociale), la diffusion massive d'une méthodologie disciplinaire ne suffit pas à accroître la légitimité de la discipline qui l'a développée. Cette diffusion est plutôt une occasion renouvelée d'interroger la légitimité de la discipline elle-même. De prime abord, le raisonnement pourrait même paraître sensé. Après tout, une fois que toutes les disciplines ont adopté les principes de *[insérer discipline à choix]*, faut-il encore une section qui fasse de *[insérer discipline à choix]* sa spécialité ? Ne s'agit-il pas là d'un dispendieux gaspillage des deniers publics ?

On le sait : la contestation du bénéfice des sciences humaines dans l'espace politique et économique est un phénomène endémique. De même que l'on doit désormais s'habituer à un COVID hivernal, on pourrait être incité à observer sans grande inquiétude les charges périodiques contre l'enseignement du grec et du latin. L'institution sait, en général, faire corps contre ces attaques extérieures. De tribunes en plateaux télé, les sciences humaines défendent leur bilan en invoquant, sources à l'appui, des externalités négligées dans les réflexions budgétaires et l'on finit toujours par trouver un directeur de banque qui a fait latin-grec pour « sauver l'honneur ». En matière de place accordée aux SHS dans le débat public, il n'y aurait donc *Nihil novi sub sole*. Au nom de quelle étrange prétention l'informatique pour les sciences humaines échapperait-elle à ces assauts politiques et, surtout, vaut-il vraiment la peine de s'y attarder ?

Pour qui est coutumier de ces attaques, il y a pourtant une différence fondamentale entre les remises en question du grec et de l'informatique

en Lettres : lorsque l'enseignement de cette dernière est remise en question, le plus souvent, l'ennemi n'est pas extérieur. L'éventuelle contestation provient toujours des organes de gouvernance qui peinent à comprendre son inscription dans un contexte de Lettres. Tout se passe comme si, pour d'obscures raisons, notre discipline était contrainte de présenter plus régulièrement ses papiers d'identité à ses collègues. On peut d'ailleurs penser (suivant en cela un phénomène bien documenté de domination sociale) que les nombreuses mutations de la section depuis sa création (qui modifie ses plans d'études plus souvent que ne le requiert l'administration et n'a pas hésité à en créer trois supplémentaires) résultent non seulement de l'innovation propre à ses membres, mais aussi d'une insécurité systémique. De même que les minorités sont davantage incitées à maîtriser les subtilités de la langue pour disposer de la flexibilité requise par les contextes de communication (Singy, 2004), notre section serait ainsi plus incitée à se réinventer.

C'est cet effet de système que la première partie du présent article aimerait mettre en perspective, en cherchant les facteurs qui expliquent cette différence de traitement.

1.2 Généalogie d'une (in)discipline

Puisqu'il n'est pas de description de système sans mémoire des états – et sans aller jusqu'à créer une chaîne de Markov de la section – prenons le temps d'un bref pas en arrière. La section que nous connaissons aujourd'hui sous le nom des « Sciences du langage et de l'information » à l'UNIL est le résultat d'une fusion, intervenue en 2013, entre deux sections indépendantes qui disposaient (et disposent encore) de plans d'études et de recherches largement dissociés. Avant cette fusion, la Faculté des lettres comptait une section de linguistique et une section d'informatique et méthodes mathématiques (IMM). L'IMM débute avec deux professeurs : Eric Keller, informaticien (spécialisé en analyse de la parole) et François Bavaud, physicien et statisticien (spécialisé, notamment, en statistique spatiale et en théorie de l'information).

Deux profils en somme très distincts, bien plus éloignés entre eux que deux collègues de Français médiéval et de Français moderne. On peut même affirmer que les deux professeurs ne sont apparentés que

par un phénomène de métacontraste, le même qui explique que deux Parisien·ne·s auront bien davantage de chances de se trouver proches dans un trek à Madagascar que sur les Champs-Élysées (voir à ce propos Salzarulo, 2006). Comment une si petite section a-t-elle pu, dès sa naissance, rassembler deux orientations aussi éloignées ? Pour le comprendre, il nous faut faire un (dernier, promis) retour en arrière.

Dans une ère presque préhistorique, bien avant la Faculté des géosciences et de l'environnement (FGSE) et son bâtiment Géopolis en damier, la géographie n'était pas une faculté, mais une section rattachée à la Faculté des lettres. Son autonomisation (voulue, selon divers témoignages, par les géologues, qui se trouvaient peu de liens avec les SHS) aurait pu déporter l'entier des personnes engagées en FGSE, mais un reliquat a subsisté en Faculté des lettres au travers d'un poste à cheval entre les deux facultés. Dès sa création, la section s'est donc trouvé traversée par une interdisciplinarité thématique (la synthèse de la parole et la statistique spatiale), une transversalité disciplinaire (l'informatique, la statistique et la géographie) et un attelage inter-institutionnel.

Nous reviendrons sur l'indiscipline induite par cette structure. Pour le moment, gardons à l'esprit que ce type d'hybridation – volontiers promue dans les plans stratégiques pluriannuels à tous les niveaux de la hiérarchie académique – est le plus souvent négligée et extradée aux confins des formations disciplinaires. Fidèle à l'idée-fiction d'une « excellence » monodisciplinaire, il conviendrait en effet *d'abord* d'atteindre l'expertise dans un champ donné (et donc d'acquérir les jeux de langage et toutes les identités sociales qui accompagnent l'habitus propre à cette discipline) pour *ensuite* s'ouvrir à une rencontre féconde avec une discipline complémentaire.

On sait pourtant qu'une telle approche revient à peu près à considérer qu'il suffirait, pour éduquer avec succès à la différence, de maintenir les gens dans l'entre-soi durant cinq ans, puis de s'ouvrir à l'autre durant un séminaire de deux semaines. Or de tels « patchs » extradisciplinaires débouchent le plus souvent sur un sentiment d'exotisme. Rafraîchissants pour les éventuels dominants, stigmatisants pour les éventuels dominés, ces ajouts de dernière minute conduisent au mieux à des regrets (« si j'avais su ! »), au pire, au renforcement d'un certain nombre de préjugés

identitaires (« ces gens ne comprennent rien à rien ») qui contribueront à hypothéquer d'autres initiatives d'hybridation des compétences. À l'instar du plurilinguisme, l'interdisciplinarité gagne à être expérimentée le plus tôt possible.

Ceci dit, puisque les autres (mono)disciplines ne perçoivent pas l'intérêt de réviser leurs propres programmes en vue d'une collaboration interdisciplinaire, que peut accomplir une section composée de forces très limitées et très hétérogènes dans un contexte où l'ensemble des pairs nourrissent une incompréhension susceptible de se muer en un sentiment d'incompatibilité, voire de rejet ? Dans le jeu de la (sur)vie académique, deux stratégies se dessinent.

1.3 Du côté de chez Conway

La première consiste à établir des collaborations nationales ou internationales afin de chercher à décloisonner la section et accroître ses effectifs. Une stratégie porteuse, qui requiert cependant une certaine masse critique du côté des professeur·e·s. Comment, sinon, courir le monde tout en siégeant au Conseil de Faculté et dans bon nombre de commissions locales influentes, nombre redoublé par les Facultés finançant un poste conjoint ? Sous cet angle, il n'est pas étonnant de constater qu'en IMM, les partenariats locaux et une importante implication dans les structures de gouvernance de la Faculté ont longtemps primé sur les consortiums internationaux. Dans ce domaine, notre collègue François Bavaud n'a pas ménagé ses heures : multiples législatures dans les deux conseils de Faculté, mandat de vice-doyen, participation active aux groupes de travail à l'origine des humanités numériques, échanges interminables avec la Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP) et la Conférence intercantonale de l'instruction publique de la Suisse Romande et du Tessin (CIIP) pour faire reconnaître une informatique plurielle, etc.

La seconde stratégie consiste à opter pour une approche de « service ». L'avantage d'une telle orientation est double. D'une part, elle permet de faire la preuve par l'acte de la pertinence des méthodologies informatiques et statistiques dans le domaine des sciences humaines et sociales. D'autre part, elle permet de collecter des données sur la

pratique scientifique de ses pairs et de faire évoluer langage, méthodes et objets au fil de ces mises en contact. On peut d'ailleurs considérer qu'une partie, au moins, des humanités numériques au niveau international émerge de démarches d'hybridation cumulées et de situations institutionnelles aussi singulières que la nôtre.

Cette seconde stratégie a été largement explorée par notre section. Le nouveau nom donné à notre discipline en 2013 (Informatique pour les sciences humaines ou ISH) et l'implication prépondérante de ses membres dans la mise sur pied de deux programmes de spécialisation, d'un Master interfacultaire en humanités numériques, de formations continues et même d'une consultation statistique en témoignent. Une stratégie assurément gagnante, qui a permis de nouveaux engagements à même de mieux distribuer les charges et d'explorer d'autres objets et d'autres méthodes.

1.4 Enfin libres ?

En se mettant « au service » de ses pairs, l'informatique « pour » les sciences humaines s'est-elle enfin guérie de ses procès en légitimité ? En partie, oui. Reste que l'entrisme méthodologique ne suffit pas à déclencher une (r)évolution des épistémologies locales. En témoignent les discussions récurrentes avec différent·e·s collègues pour qui la création logicielle, toute utile qu'elle soit dans les cas particuliers (souvent leurs propres projets de recherche), ne devrait en aucun cas faire partie des objectifs de formation académiques d'une Faculté des lettres. En bref, si on demande moins fréquemment aux fans *des chiffres et des lettres* de présenter leurs papiers d'identité, force est de constater que l'on est encore loin de pouvoir se promener sans porte-monnaie. Comment expliquer ce double standard ?

1.5 Une certaine conception de la pratique

La doxa en SHS (bien plus que les textes réglementaires) tend à réserver le développement de compétences pratiques aux HES. Un discours abondamment repris par plusieurs disciplines qui ont été jusqu'à créer leur identité sur une ferme opposition à la pratique au profit d'une approche essentiellement critique. En somme, que des collègues à l'interne puissent « donner un coup de main » pour se mettre en conformité

avec le volet numérique des exigences des principaux bailleurs de fonds est bienvenu, mais de là à enseigner le développement logiciel, voilà qui semble (à encore trop de collègues) inopportun, voire déplacé.

Il suffit pourtant de suivre une semaine de cours dans n'importe quelle discipline pour découvrir que la pratique est omniprésente en SHS. Ainsi, l'histoire et l'esthétique du cinéma accorde une place cruciale à l'analyse de séquence dans ses plans d'études et quiconque a effectué des analyses littéraires dans n'importe quelle discipline associée aux langues et littératures a dû faire face à des formalismes qui feraient fuir bien des étudiant·e·s terrorisé·e·s par la programmation et les mathématiques. Lorsqu'un·e collègue analyse un poème, il ou elle va puiser dans un formalisme très informé qui, s'il tait son nom, présente des affinités évidentes avec le type d'analyses propres à l'informatique textuelle.

Tout se passe donc comme si, dans les SHS universitaires, la pratique, bien que très présente, devait être tue. Quiconque a vécu l'intégration des objectifs de formation dans les plans d'études de notre Faculté se souvient de « discussions » plus que houleuses. Il faut bien le reconnaître : qu'on le veuille ou non, trop souvent en SHS, quiconque s'obstine à mettre un trop fort accent sur les méthodes s'expose à être taxé·e d'imbécile, à l'instar de qui se concentre sur le doigt quand le sage montre les étoiles.

Les contributions significatives de Callon et Latour sur l'acteur-réseau (Callon & Latour, 1981) suffisent à renvoyer l'imbécilité du côté du « sage » qui s'obstine à négliger le doigt. Mais plutôt que de retourner le compliment, prenons ce reproche au sérieux : et si l'informatique se trouvait constamment en procès de légitimité parce qu'elle se focalise sur le doigt au détriment des étoiles ?

1.6 Allô Allô Monsieur l'ordinateur

Poser cette question revient à se demander si notre dispositif privilégié (l'ordinateur) ne serait pas la source d'un malentendu durable. Les SHS ayant pour priorité de travailler sur l'humain et ses productions au travers de l'intuition des humains eux-mêmes, l'instrument informatique pourrait être perçu comme un facteur d'aliénation conduisant à une

corruption des objets et des méthodes propres aux SHS. En somme, puisque « *de te fabula narratur* » (Horace, Satires, 1, 64-79), il faudrait que « *Cor cordi loquitur* » (St-François de Salles). Dans pareil contexte, on pourrait comprendre que l'ordinateur soit perçu comme l'instrument d'une violence faite à l'intuition, le dispositif informatique passant pour tout sauf « humain ».

Cette hypothèse semble d'autant plus solide que notre discipline n'est pas la seule à souffrir de cette mise à l'écart. De même que l'injection d'une dimension ludique dans un objet culturel a pour effet de dévaluer son prestige (les livres-jeux n'auront jamais la légitimité de la littérature blanche), l'art génératif a, dans son versant informatisé, longtemps souffert du même type de discrédit. Et tout indique qu'il faudra encore bien des Vera Molnar pour convaincre le milieu artistique, la pensée populaire et certains milieux académiques de la capacité des ordinateurs à stimuler et affiner notre intuition.

À la décharge du public et des milieux autorisés, il faut bien admettre que les grandes firmes technologiques n'aident pas à penser l'ordinateur dans une continuité culturelle. L'obsolescence des dispositifs – mise en scène sous la forme d'une innovation des petits pas favorable à la rythmique financiarisée des grands acteurs industriels – fait même volontairement obstacle à cette inscription historique. Sous cette lumière, notre dernier ordinateur/smartphone n'est pas le plus récent avatar des travaux d'Ada Lovelace sur la machine de Babbage : c'est un ovni techno-magique venu accroître notre créativité et notre productivité. Rien d'étonnant, donc, à ce que l'on associe moins l'informaticien·ne à un·e prolongation instrumentée d'un·e philosophe, qu'à une sorte d'Inspecteur Gadget un peu étourdi sur les vrais enjeux d'une recherche en SHS.

Or, pour qui connaît l'histoire des techniques, cette lecture est franchement bizarre. On a beau avoir tendance à l'oublier, durant l'essentiel de leur histoire, les livres n'ont pas été si différents des ordinateurs en matière de technicité. Réservés à une caste d'initiés, le *codex* et ses descendants ont été longtemps l'apanage d'élites puissantes, seules capable de les déchiffrer. Quant aux bibliothèques qui cherchent aujourd'hui à nous ramener vers leurs collections par tous les moyens, elles ont été

des lieux au moins aussi interdits que les bureaux de la défense ou que les laboratoires académiques accueillant l'ENIAC et ses héritiers.

Cette communauté de destin ne doit évidemment rien au hasard. En plus d'être les supports de connaissance par excellence, livres et ordinateurs proposent des manières complémentaires d'altérer notre rapport au temps : tandis que la lecture permet d'accéder au passé et au futur (et donc de dialoguer avec les morts et les encore non vivant·e·s), les ordinateurs permettent en sus d'accélérer le temps (et donc de ralentir nos vies) en réalisant une myriade d'opérations à la seconde.

On l'a compris : ordinateurs et livres sont donc bien plus apparentés que ne le laissent supposer nos imaginaires habitués à opposer le caractère chaleureux du papier à la froideur du silicium. Reste que les acteurs industriels ne sont pas seuls responsables des attaques en légitimité adressées à une informatique désireuse de se déployer dans un contexte de Faculté des lettres. Si nous avons été accusés (à tort) de nous focaliser sur le doigt, nous sommes bel et bien coupables d'une chose : ne pas avoir désigné nos propres étoiles.

1.7 To the stars and beyond

Qui, dans ses cours d'histoire de l'art au gymnase, a entendu parler de Vera Molnar ? Combien d'heures ont été consacrées à l'émergence de l'informatique dans les cours d'histoire ? Et combien à l'importance de la logique philosophique (Bertrand Russell, Bolzano) et aux questions de calculabilité ? À moins d'un hasard qui mériterait d'être documenté, il faut bien reconnaître que l'ordinateur en tant qu'objet culturel est (et demeure) le grand absent de l'ensemble des disciplines qui conduisent à l'enseignement tertiaire en sciences humaines et sociales. C'est dire si les dispositifs qui conditionnent désormais l'ensemble de nos rapports au monde s'inscrivent dans un vide historique et conceptuel frappant.

Tout indique hélas que les différentes initiatives en éducation numérique au secondaire 1 et 2 ne parviendront pas seules à combler ce vide. La plupart du temps peu dotés en heures d'enseignement et délégués à des enseignant·e·s prioritairement formé·e·s et invité·e·s à regarder le doigt (dans le bon sens du terme), les plans d'études en informatique dans le primaire comme dans le secondaire mettent, sans surprise, la

focale de l'apprentissage sur « la différence qui fait la différence » (Bateson, 1972), à savoir la programmation. Pas le temps, en somme, de narrer la fascinante histoire de l'informatique ou de poser un regard de *critical code studies* sur la programmation elle-même (Pante, 2021). Tout indique donc que c'est à nous, enseignant·e·s du tertiaire, qu'il revient d'explorer cette voie.

Il est en effet grand temps que notre discipline passe de l'informatique *pour* les sciences humaines à l'informatique *en tant que* science humaine. Ce changement de paradigme ne consiste pas à renoncer à la versatilité et à l'agnosticisme de nos méthodes, mais à les compléter à double titre. D'une part, en situant les méthodes dans leur contexte socio-historique (ce que les humanités numériques font volontiers) et en se dotant, à l'instar des autres disciplines, d'objets qui nous soient propres.

Le premier de ces objets potentiel tombe sous le sens : il est urgent de faire l'histoire de l'ordinateur, non seulement en tant que calculatrice capable de résoudre des opérations mathématiques, mais aussi en tant que dispositif d'évolution de la pensée humaine situé dans un contexte historique et s'inscrivant dans une tradition philosophique et politique propre. On sait que si la Suisse a été l'un des berceaux de bon nombre d'innovations informatiques, l'histoire de ce patrimoine menacé reste à écrire de toute urgence.

Or faire l'histoire de ces objets nous mène directement aux productions qu'il autorise, à commencer par le jeu vidéo. Nous tenons ici notre second objet, voire le premier si l'on prend en compte notre contexte institutionnel. Souvenons-nous que, en Faculté des lettres, la plupart de nos collègues ont construit leur identité disciplinaire sur des œuvres artistiques ou culturelles. Pourtant, si le roman, la peinture, la sculpture, le film, la photographie et la musique ont su créer leur discipline, le jeu vidéo – qui agrège pourtant bien souvent la totalité de ces formes d'expression culturelles – n'a été pendant longtemps abordé que le temps d'une escapade exotique souvent initié·e par un·e étudiant·e passionné·e et un·e enseignant·e curieux ou curieuse (dans tous les sens du terme).

On peut penser que cette omission n'est que le symptôme de la lente évolution des mentalités générales ainsi que de l'inertie (supposée)

du monde académique face à l'intégration des objets de la culture populaire. Pourtant, les contre-exemples abondent : la bande dessinée et le théâtre sont traités. Pourquoi pas le jeu (vidéo) ? Osons une réponse qui dépasse les paniques morales et les préjugés disciplinaires : parce que la complexité de l'objet requiert un cumul de compétences qui ne se limite pas au cœur de cible des méthodes en SHS, mais demande une hybridation des parcours de formation et des identités disciplinaires.

Pour mieux le comprendre, prenons appui sur une discipline voisine. Qu'attend-on d'une bonne analyse d'un tableau ? Non seulement qu'elle nous renseigne sur la personne qui l'a produit et le courant auquel il appartient, mais aussi que nous soit précisée son inscription dans les méthodes de son temps en soulignant les contraintes posées par les matériaux de l'époque (pigments disponibles, etc.) Une bonne analyse de jeu (vidéo) ne saurait se contenter de moins. Or qui mieux qu'une section d'informatique en Faculté des lettres pourrait aller, en pleine compétence, à la rencontre de ces objets complexes qui mêlent la narration, l'image, les mathématiques (notamment probabilistes) et des stratégies computationnelles avancées ? Comment pourrait-on, en effet, sans disposer de compétences croisées en sciences humaines et en ingénierie, corréler une production vidéoludique aux dispositifs qui ont rendu possible son émergence ?

Loin de se confiner à l'informatique *pour* les sciences humaines, l'informatique *en tant que* science humaine pourrait donc être repensée comme une discipline mobilisant la modélisation et les techniques computationnelles pour étudier les productions artistiques de l'informatique à commencer par les jeux (vidéo), entendus comme des objets pluridimensionnels demandant (a minima) un double regard. Car oui, l'ensemble des outils computationnels propres à l'informatique et aux méthodes mathématiques peuvent et souvent doivent être mobilisés, que ce soit pour analyser les jeux existants ou pour accompagner leur développement.

Depuis quelques années, notre section s'est engagée dans cette voie en accordant plus de place, dans ses plans d'études, dans ses projets de recherche comme dans ses recrutements, à ces objets si populaires et pourtant dédaignés par les autres disciplines. Nous allons à présent

illustrer la fécondité de cette approche en montrant comment la théorie des graphes et l'analyse de réseaux peuvent être appliqués à des objets littéraires qui se passent volontiers de tout dispositif numérique : les *Livres dont vous êtes le héros* (LDVELH).

2 Études de cas - La fiction interactive

2.1 Le jeu sous toutes ses formes

Si le jeu vidéo a été pointé comme un objet privilégié et légitime pour notre discipline, il n'est évidemment pas le seul artefact ludique dont puisse s'emparer une approche computationnelle. Des formes hybrides, souvent reléguées à la marge des disciplines qui devraient pourtant les prendre pour objet, méritent également toute notre attention. En 1997 déjà, Espen Aarseth, narratologue indiscipliné, déplorait que le corpus de son champ fasse l'impasse sur les « cybertextes », soit ceux capables non seulement de manipuler leur lectorat (objectif de toute littérature), mais aussi de se manipuler eux-mêmes, que ce soit de manière automatique ou avec l'aide d'un « opérateur » (Aarseth, 1997).

On pourrait attendre d'un narratologue intéressé par la littérature qui bifurque qu'il se limite aux *Cent mille millards de poèmes* (Queneau, 1961) et aux explorations qui vont de *Marelle* (Cortázar, 1963) à *La maison des Feuilles* (Danielewski, 2000) en passant par les expérimentations de l'OuLiPo. Aarseth invite pourtant à dépasser le livre et inclut dans la cybertextualité des objets aussi divers que le Yi-Jing, les jeux vidéo, internet et le tarot de Marseille. L'enjeu de cette classification est considérable : en contestant des regroupements orientés sur les supports et les dispositifs (les livres avec les livres et les jeux avec les jeux), Aarseth invite à créer une nouvelle taxonomie qui fasse primer les opérations (la fameuse agentivité chère aux formes ludiques) sur les supports afin, notamment, de renouveler l'épistémologie d'une discipline selon lui bien trop repliée sur ses œuvres canoniques.

Près de trente ans plus tard, force est de constater que cet appel n'a pas été entendu : dans l'enseignement des langues et de la littérature, les cybertextes conservent un statut exotique (les fameuses deux semaines d'interdisciplinarité) et les chances de rencontrer, dans l'entier d'un parcours littéraire, la fiction interactive informatisée ou les célèbres

LDVELH demeurent extrêmement faibles. Au travers de formations continues qui demeurent pour l'essentiel le fait d'initiatives individuelles, certain·e·s enseignant·e·s travaillent pourtant à renouveler leurs pratiques, en cherchant à donner un visage aux objectifs numériques proposés par les différentes révisions des plans d'études romands. Le solide intérêt pour les *écritures numériques* et ses dispositifs constitue à ce titre une opportunité d'inviter ce public à repenser les limites de leur corpus.

Tout indique cependant qu'il appartient aux « humanistes numériques », entendus ici comme des personnes « dotées d'une formation aux arts et aux techniques computationnelles » de s'emparer de ces objets hybrides et d'en rendre compte avec des outils, des méthodes et une largeur de vue qui refuse de les mutiler.

2.2 Enjeux d'une analyse des LDVELH

Depuis six ans, mes étudiant·e·s et moi encodons à la main des livres-jeux afin de les étudier grâce aux instruments de l'analyse de réseau. Dans un contexte d'enseignement de l'informatique *en tant que* science humaine, l'étude des LDVELH permet évidemment de visibiliser l'architecture d'états du monde souvent invisibilisés par les jeux vidéo et d'éprouver, au travers d'un dispositif papier, les contraintes bien réelles d'une pensée computationnelle. L'analyse de cette littérature interactive est ainsi l'occasion de découvrir que, là où trois variables informatiques permettraient aisément de stocker une combinaison de clefs nécessaires à l'ouverture d'un coffre, le dispositif papier doit créer une structure combinatoire complexe, bien visible sur la figure 1, tirée du fameux coffre final du *Sorcier de la montagne de feu* (Jackson & Livingstone, 1982). Ce type d'approche permet donc d'exposer une composante combinatoire et computationnelle en l'absence de tout dispositif digital et de réaliser le geste souhaité par Aarseth en floutant les ontologies usuellement fondées sur les supports au profit d'un continuum d'expériences.

En plus de visibiliser diverses contraintes mécaniques, ce va-et-vient entre lecture proche et distante (*close & distant reading*) permet aussi de quantifier l'importance des différents types de nœuds (combat, test

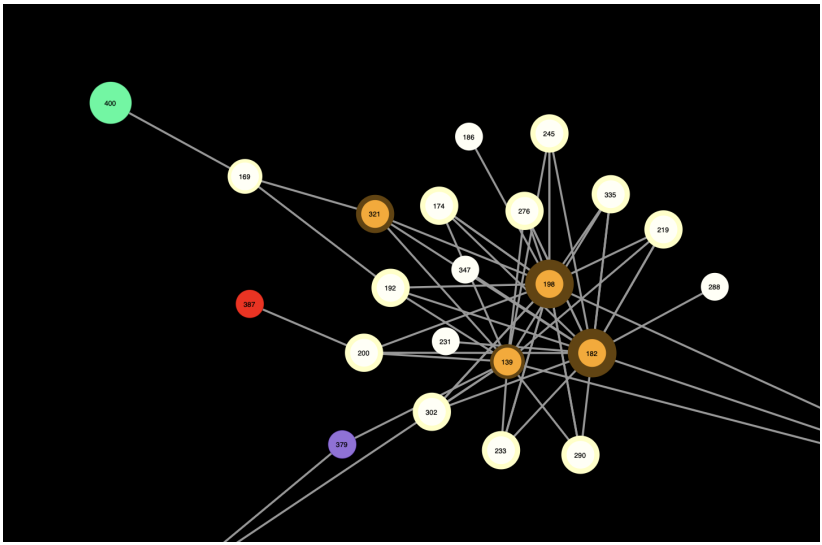


FIGURE 1 – Machine textuelle permettant l'insertion de trois clefs dans un ordre aléatoire pour déverrouiller le coffre final du *Warlock of the Tabletop Mountain* (Jackson & Livingstone, 1982).

de compétence, objet requis, etc.) et de faire émerger des patterns narratifs en les rapportant aux indicateurs usuels de l'analyse de réseau (densité, diamètre, hubs, cliques de nœuds, etc.). Pour peu que l'on dépasse la pure description, cette mise en relation offre une occasion unique d'étudier les ressentis et les leviers d'une agentivité qui peut (ou non) accompagner la lecture de ces textes qui prétendent – souvent de manière abusive – nous mettre au centre de l'aventure et valoriser nos choix (voir figure 2).

Du point de vue de la recherche, l'un des bénéfices majeurs d'une approche qui choisit de mettre l'accent sur le ressenti d'agentivité est qu'elle rend impossible une mutilation de l'objet à l'étude. Si le « sentiment » de liberté n'est accessible qu'au travers d'une expérience de première main du livre-jeu analysé via une lecture attentive, la pertinence de ce sentiment (notre fameuse intuition) ne peut être évaluée qu'une fois rapportée à une étude précise et quantifiée des mécaniques ludiques qui autorisent, simulent ou contredisent cette agentivité.

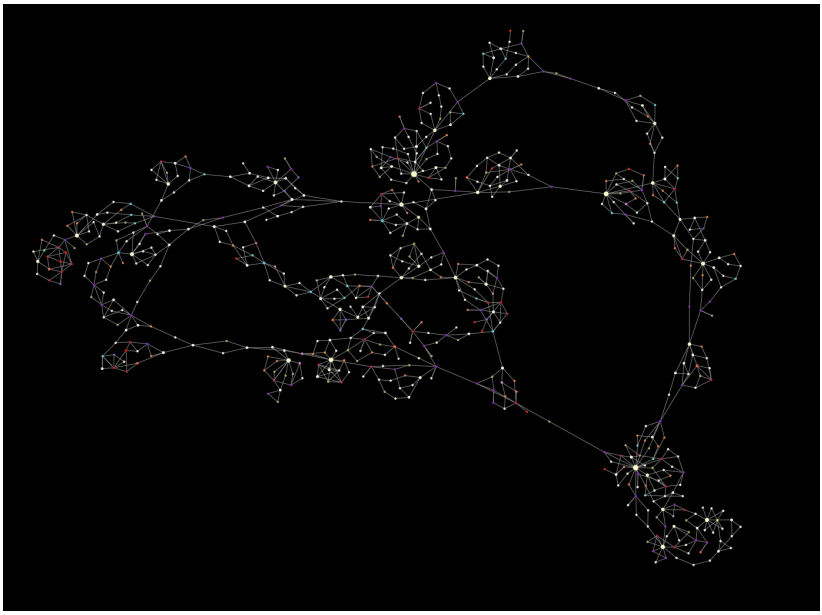


FIGURE 2 – Réseau des 619 passages de *Bloodfeud of Altheus* (John Butterfield, 1985).

Ces interrogations fécondes au niveau d'un LDVELH donné gagnent évidemment à être élargies dans une approche de littérature ludique comparée prenant pour variables les autrices et auteurs, les époques de production, et la prétention desdites œuvres au prestige littéraire. En élargissant la focale de la sorte, il devient possible de tester des hypothèses sur l'impact du voisinage intermédial et de formuler des questions de recherche aussi stimulantes que pertinentes pour toutes les disciplines de Lettres. On peut ainsi se demander si la propagation des jeux vidéo a impacté la composante mécanique des livres-jeux, si les mécaniques ludiques survivent à une prétention au prestige littéraire, etc.

Mais par-delà les études combinatoires et les questions de recherche, cette matière offre également une occasion exceptionnelle de dresser l'histoire d'un genre méconnu et trop souvent associé à des représentations qui font violence à sa diversité réelle. Car si le nom canonique

de LDVELH est devenu le porte-étendard de toute une forme littéraire en prescrivant un lectorat masculin (« H » pour « héros » et non « héroïne »), il invisibilise aussi les sources réelles du genre. Apprendre aux étudiant·e·s que cette littérature qui bifurque est née en 1930 sous la plume de deux femmes avec *Consider the Consequences !* (Webster & Hopkins, 1930) ne fait pas que fournir un contexte littéraire aux objets étudiés : ce renvoi à l'histoire intègre aussi, au sein même de nos objets de recherche, des *role models* dont on connaît l'importance dans une discipline aussi genrée que la nôtre.

2.3 De la créativité computationnelle à la création littéraire

L'envie est dès lors forte, pour chaque étudiant·e formé·e à ces différentes formes littéraires de s'essayer à l'écriture. Et autant le dire d'emblée : quels que soient les préjugés précités portant sur la place de la pratique en SHS, il serait absurde et contre-productif de se limiter à la seule analyse des œuvres. Cette posture purement analytique est en effet intenable, non seulement parce que l'essentiel des logiciels destinés à la création de fictions interactives ont vu leur coût d'apprentissage diminuer drastiquement, mais aussi parce que nos étudiant·e·s, par la diversité de leurs profils, sont susceptibles de devenir des actrices et acteurs privilégiés d'un genre qui a grand besoin de leurs imaginaires. Le titre du manifeste d'Anna Anthropy (2012) *Rise of the videogame zinesters : how freaks, normals, amateurs, artists, dreamers, drop-outs, queers, housewives, and people like you are taking back an art form* montre bien que, par-delà la simple diffusion d'une technique, la fiction interactive peut constituer (et constitue d'ores-et-déjà) l'un des lieux privilégiés d'un *empowerment* culturel des minorités. À l'heure d'une libération croissante de la parole des femmes et de la communauté queer, une section d'informatique gagne à se doter de ses propres sessions de *creative writing*, par ailleurs de mieux en mieux acceptées dans les contextes académiques francophones.

Rappelons qu'un tel glissement vers une créativité artistique prenant appui sur des démarches computationnelles ne va pas nécessairement de soi, l'informatique étant (dans les imaginaires collectifs) plus volontiers associée à une dimension de calcul. Pourtant, entre *Twine*, *Bitsy*, *RenPy*

et *Kaplay.js*, rien n'empêche désormais de proposer à nos étudiant·e·s de mobiliser leurs compétences disciplinaires en SHS en contribuant à une recherche-création qui ne craint plus d'explorer d'autres voies. Les *Swiss Game Awards* décernés à nos étudiant·e·s par la *Swiss Game Developer Association* en 2022 (*Vote Now !*, Joël Rimaz) et en 2023 (*Fix it !*, Amina Matt & Mélissa Matti) confirment l'intérêt croissant de la cité pour un savoir qui ose se mettre en jeu.

3 Ce n'est qu'un au revoir

Souhaitons que ce trop bref panorama des usages des graphes et des réseaux dans le cadre des LDVELH contribue à illustrer à quel point les aspects computationnels, ludiques et littéraires cohabitent au sein de bon nombre de nos objets culturels, numériques ou non. L'expérience fructueuse de ces trente dernières années, tant au sein de notre section qu'au sein du Gamelab Lausanne (qui soufflera ses dix bougies en 2026) a démontré que croiser les disciplines n'est pas seulement l'occasion de créer de l'interdisciplinarité, mais aussi de favoriser de l'indiscipline, dont on connaît, à l'instar de l'erreur, l'importance cruciale pour toute science qui ne cesse d'interroger ses dogmes et se fait forte d'explorer de féconds chemins de traverse.

Gardons ceci à l'esprit : l'informatique *en tant* que science humaine gagne non seulement à promouvoir ses méthodes, mais aussi à poursuivre et à revendiquer l'étude des objets culturels indissociables de son ADN. En bref, à ne sacrifier ni le doigt, ni les étoiles.

Références

Aarseth, E. J. (1997). *Cybertext: perspectives on ergodic literature*. Johns Hopkins University Press, Johns Hopkins University Press, Baltimore, USA.

Anthropy, A. (2012). *Rise of the videogame zinesters: how freaks, normals, amateurs, artists, dreamers, drop-outs, queers, housewives, and people like you are taking back an art form*. Seven Stories Press, New York, USA.

Bateson, G. (1972). *Steps to an ecology of mind*. University of Chicago Press, Chicago, Ill.

Callon, M. & Latour, B. (1981). Unscrewing the big Leviathan: How actors macro-structure reality and how sociologists help them to do so. In Knorr-Cetina, K. & Cicourel, A. (éd.), *Advances in social theory and methodology: toward an integration of micro- and macro-sociologies*. Routledge & Kegan Paul, London, UK.

Cortázar, J. (1963). *Rayuela*. Editorial Sudamericana, Buenos Aires, Argentine.

Danielewski, M. Z. (2000). *House of leaves*. Pantheon Books, New York, USA.

Jackson, S. & Livingstone, I. (1982). *The warlock of Firetop Mountain*. Puffin Books, Harmondsworth, UK.

John Butterfield, David Honigmann, P. P. (1985). *Bloodfeud of Altheus*. Puffin Books, Harmondsworth, UK.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, 1:54–68. London, UK.

Pante, I. (2021). De l'édition numérique à l'édition du numérique. *Bulletin de l'Académie suisse des sciences humaines et sociales*, 27(3):36–39. Berne, Suisse.

Queneau, R. (1961). *Cent mille milliards de poèmes*. Gallimard, Paris, France.

Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation*, 9(1):13. Guildford, UK.

Singy, P. (2004). *Identités de genre, identités de classe et insécurité linguistique*. L'Harmattan, Paris, France.

Webster, D. & Hopkins, M. A. (1930). *Consider the consequences!* The Century Company, New York, USA.

Cahiers du CLSL, n° 69, 2025, pp. 207-218

Les humanités numériques – vers une « mathématique originale » ?

Michael Piotrowski
Université de Lausanne
michael.piotrowski@unil.ch

Résumé

Cet essai explore le potentiel de la pensée formelle et des mathématiques en tant qu'outils fondamentaux dans les sciences humaines, comme l'envisageait Granger, qui soulignait leur rôle « non pas seulement comme réduction des phénomènes aux calculs, mais aussi comme invention de structures nouvelles, voire même d'une mathématique originale ». D'un point de vue historique, les mathématiques sont passées de l'art de la mesure à une « science des modèles ». C'est là qu'il faut chercher la « mathématique originale » pour constituer la base conceptuelle des humanités numériques, qui mettent l'accent sur la modélisation formelle plutôt que sur la simple automatisation des calculs. Cette perspective appelle à une intégration plus poussée des méthodes informatiques avec leurs fondements logiques et mathématiques afin de faire progresser les cadres épistémologiques des sciences humaines.

1 Introduction

Le problème de la définition des humanités numériques n'est pas tant le fait qu'il y ait tant de tentatives de définitions contestées, mais plutôt le fait que de nombreuses personnes dans le domaine, rejettent toute définition, affirmant que les humanités numériques sont en quelque sorte « indéfinissables ».

Dans « Décomposer les humanités numériques » (Piotrowski & Xanthos, 2020), nous avons soutenu qu'une définition est nécessaire pour des raisons scientifiques, institutionnelles et politiques. Je ne reviendrai

pas sur ces arguments ici. Ce qui est important, c'est que nous fassions une distinction entre les humanités numériques *appliquées* et *théoriques* et que nous donnions des explications concises pour ces termes :

Humanités numériques appliquées. Ce terme désigne les domaines de recherche qui, comme l'histoire computationnelle ou les études littéraires computationnelles, s'inscrivent dans une discipline des sciences humaines et ont pour objet la construction de modèles formels des phénomènes étudiés par cette « discipline mère ». La différence entre les études « traditionnelles » et « computationnelles » porte donc spécifiquement sur la nature des modèles qu'elles visent à construire : dans le cas de ces dernières, il s'agit de modèles *formels* qui peuvent être manipulés par des ordinateurs. Pour le reste, ils partagent les objets de recherche et les objectifs des disciplines des sciences humaines auxquelles ils appartiennent.

Humanités numériques théoriques. Elles étudient les propriétés générales de ces modèles à un niveau d'abstraction plus élevé. En d'autres termes, les humanités numériques théoriques créent et étudient les *métamodèles* dont l'application concrète dans les sciences humaines est le domaine des humanités numériques appliquées, ainsi que la méthodologie de construction de ces métamodèles. Ces métamodèles prennent généralement la forme d'algorithmes et de structures de données, c'est-à-dire de programmes informatiques (Wirth, 1976). Bien qu'ils soient conçus pour un domaine d'application particulier, la question de recherche qui les sous-tend est l'*adéquation* de ces modèles, et non leur application. On peut donc dire que les humanités numériques théoriques servent de métascience pour les humanités numériques appliquées.

Cette distinction est cruciale, car les humanités numériques appliquées et théoriques ont des objets et des objectifs de recherche différents : dans le premier cas, elles appartiennent aux sciences humaines, dans le second, à l'informatique.

Maintenant, à quoi ressemblent ces modèles formels ? Comme le souligne Granger (1967, p. 19), il y a une « confusion que l'on favorise

d'ordinaire entre la pensée formelle et l'œuvre des mathématiciens. » Granger note que s'il est « bien vrai en un sens que tout formalisme scientifique efficace tend vers un statut mathématique, ce n'est pas pour autant qu'il se réduise infailliblement aux instruments *usuels* et *actuels* des géomètres. » Il se donne la tâche de « montrer la pensée formelle à l'œuvre dans les sciences humaines, non pas seulement comme réduction des phénomènes aux calculs, mais aussi comme invention de structures nouvelles, voire même d'une mathématique originale. »

2 Contexte historique et philosophique

Que pourrait donc être cette « mathématique originale » ? Pour répondre à cette question, il faut remonter dans l'histoire des mathématiques. Les racines des mathématiques en géométrie étaient, comme leur nom l'indique, concernées par la « mesure de la terre », et en particulier des champs le long du Nil après la crue annuelle. Pendant longtemps, les mathématiques sont restées associées à la mesure. Speiser (1952, p. 11) remarque que « Goethe verstand unter der Mathematik die Meßkunst », et cette identification des mathématiques aux instruments de la physique expérimentale était encore d'actualité dans la philosophie du XIX^e^ siècle; considérez, par exemple, la définition des mathématiques d'Auguste Comte (1892, p. 106) :

> Nous sommes donc parvenus maintenant à définir avec exactitude la science mathématique, en lui assignant pour but, la mesure *indirecte* des grandeurs, et disant qu'on s'y propose constamment de *déterminer les grandeurs les unes par les autres, d'après les relations précises qui existent entre elles*.

Néanmoins, il s'agit déjà d'un point de vue beaucoup plus abstrait, puisque Comte rejette explicitement la définition des mathématiques comme « la science des grandeurs » ou « la science qui a pour but la mesure des grandeurs » (pp. 97-98); même s'il parle encore de « mesure », ce sont les *relations* qu'il place au centre des mathématiques.

Dans *Les principes des mathématiques*, une série d'articles inspirés par *The principles of mathematics* de Bertrand Russell, Louis Couturat (1904, pp. 19-20) note que jusqu'au milieu du XIX^e^ siècle, la logique et

les mathématiques étaient deux domaines absolument distincts et même séparés :

> Cet état de choses a complètement changé pendant la seconde moitié du XIX^e^ siècle. D'une part, les mathématiciens furent pris de scrupules logiques inconnus de leurs prédécesseurs; ils se mirent à analyser leurs méthodes de démonstration, à vérifier l'enchaînement de leurs théorèmes, à rechercher les hypothèses ou postulats qui se glissaient subrepticement dans leurs raisonnements, enfin à dégager les principes ou axiomes d'où partaient leurs déductions et d'où dépendaient toutes leurs théories. [...] Enfin, en creusant pour ainsi dire les fondations de leur science, et en reprenant tout l'édifice en sous-œuvre, les mathématiciens furent amenés à constituer deux théories nouvelles qui devaient désormais servir de base à toutes les autres : la théorie des ensembles et la théorie des groupes; autrement dit, la science des multiplicités et la science de l'ordre. Ainsi il apparaissait que les sciences du nombre et de la grandeur n'étaient pas primitives, mais reposaient sur des doctrines d'un caractère plutôt logique que mathématique, et sur des notions qui n'avaient plus rien de *quantitatif*.

Speiser (1952, p. 12) (publié à l'origine en 1932) est alors très clair : « Die Mathematik ist ihrem Wesen nach keine Meßkunst », c'est-à-dire les mathématiques ne sont pas, par essence, un art de la mesure.

Granger (1967, p. 143) fait référence, dans une note de bas de page, à l'article de John Kemeny (1959) intitulé *Mathematics Without Numbers*. Kemeny se place explicitement dans ce développement historique en commençant l'article par la remarque qu'il y a cent ans, un mathématicien aurait défini les mathématiques comme « l'étude des nombres et de l'espace ». Il note qu'une telle définition est cependant beaucoup trop étroite pour inclure les branches les plus récentes des mathématiques modernes (p. 577). Pour nous, il est particulièrement intéressant qu'il montre les façons dont les modèles mathématiques peuvent être utilisés en relation avec des problèmes typiques des sciences sociales,

où « sciences sociales » doit être compris comme « sciences humaines et sociales ». Il note que ces sciences peuvent être caractérisées par le fait que, dans la plupart de leurs problèmes, les mesures numériques semblent absentes et que les considérations d'espace ne sont pas pertinentes (p. 577).

Lynn Arthur Steen (1988, p. 616) franchit alors l'étape logique suivante; plutôt que de considérer les « mathématiques sans nombres » comme *une* partie des mathématiques, il généralise la notion de mathématiques en tant que « the science of patterns » et relie explicitement ce développement à l'informatique, le système étroitement couplé formé des domaines d'application, des ordinateurs et de mathématiques produisant « des résultats jamais possibles auparavant et des idées jamais imaginées auparavant » (Steen, 1988, p. 612).

3 Vers une « mathématique originale » pour les humanités numériques

Puisque Kemeny (1959) discute spécifiquement d'exemples provenant des « sciences sociales », il est utile d'examiner ces exemples du point de vue des humanités numériques. Encore une fois, le sens du terme « sciences sociales » n'est pas le même en anglais et en français; dans ce contexte, l'utilisation du terme par Kemeny doit être comprise comme faisant référence aux sciences humaines et sociales.

Kemeny note d'emblée qu'il existe des problèmes d'interprétation interdisciplinaire[1] : alors que les chercheurs en sciences sociales trouvent souvent les mathématiciens incapables de les éclairer sur le modèle particulier qui les intéresse, de nombreux mathématiciens ont l'impression que les problèmes mathématiques en sciences sociales sont tout à fait triviaux. Kemeny souligne toutefois que cette dernière idée fausse est due au fait que les problèmes des sciences humaines et sociales sont trop *difficiles* pour les mathématiques actuelles, et c'est parce que les problèmes posés par les sciences humaines et sociales deviennent

1 Comme le dit Abraham Moles (1995, p. 159), « la multidisciplinarité n'existe réellement qu'à l'intérieur du cerveau d'un même individu qui n'a plus à se heurter à l'univers du faux sens ».

rapidement difficiles que seuls quelques-uns des problèmes mathématiques les plus simples ont été résolus jusqu'à présent.

Selon Kemeny, il existe essentiellement deux façons différentes de construire un modèle mathématique pour un problème qui n'implique pas de nombres ou d'espace. La première approche consiste à utiliser une branche des mathématiques qui n'emploie pas de nombres et ne traite pas de l'espace. La seconde approche consiste à introduire des nombres « par une méthode plus ou moins arbitraire », là où aucun nombre n'était apparent au départ; il peut alors être possible de former un modèle numérique d'un problème non numérique (p. 579). Kemeny donne quatre exemples de tels modèles mathématiques et les domaines mathématiques correspondants. Le premier est la théorie des graphes, et l'exemple de Kemeny est un modèle d'équilibre structurel dans les groupes sociaux. Le deuxième modèle de Kemeny utilise la *théorie des groupes*, en particulier les groupes de transformations. Son exemple est l'étude des règles de mariage dans les sociétés primitives. Son troisième exemple concerne les réseaux de communication, une application qui est d'ailleurs courante dans les humanités numériques. Afin de démontrer l'introduction artificielle de nombres, il discute du réseau en termes de matrices plutôt que de graphes. Des travaux récents en sciences humaines numériques démontrent la complémentarité des représentations graphiques et matricielles des réseaux pour différentes applications (Bavaud & Métrailler, 2023; Egloff & Bavaud, 2018). Le dernier exemple de Kemeny exploite l'utilisation d'une approche géométrique pour un problème non spatial, à savoir celui du classement, par exemple, des préférences. Kemeny montre que ce problème peut être réduit à un problème analogue aux problèmes classiques de statistiques (variables aléatoires) si l'on est capable d'introduire une mesure de distance entre les classements.

Alors que Kemeny vise à illustrer les diverses façons dont les mathématiques peuvent être utiles dans les problèmes non numériques et non spatiaux des sciences humaines et sociales, Steen (1988) affirme : « Mathematics is the science of patterns » (p. 616) :

> The mathematician seeks patterns in number, in space, in science, in computers, and in imagination. Mathematical theories explain the relations among patterns; functions and maps, operators and morphisms bind one type of pattern to another to yield lasting mathematical structures. Applications of mathematics use these patterns to "explain" and predict natural phenomena that fit the patterns.

Steen ne mentionne pas les applications dans les sciences humaines et sociales, mais sa perspective sur les mathématiques rend finalement évidente la relation entre les motifs artistiques (au sens le plus large) et la pensée mathématique que Speiser (1952) n'a pas pu exprimer, raison pour laquelle il a ouvert son livre par ces mots « Das Wesen des mathematischen Denkens unmittelbar in Worten zu beschreiben ist nicht möglich » (Speiser, 1952, p. 11).

4 Implications pour les humanités numériques

L'évolution historique des mathématiques décrite dans la section précédente est également décisive pour le développement d'une « mathématique originale » comme base des humanités numériques, que je crois nécessaire pour aller au-delà du simple « angle opératoire », comme le dit Mario Borillo (1985, p. 5); il note qu'il y a eu, dès le début de l'utilisation des ordinateurs dans les humanités et les sciences sociales, deux conceptions différentes concernant la nature des relations avec l'informatique :

> La première, qui est encore la plus courante, les envisage surtout sous un angle opératoire et il est vrai que l'ordinateur est un instrument capable de modifier profondément les conditions matérielles dans lesquelles s'effectue la recherche dans les sciences de l'homme. L'autre point de vue lie cette dimension technique à son socle conceptuel (logico-mathématique) et voit dans le recours à l'informatique un facteur susceptible de faire évoluer également les cadres théoriques et les référents épistémologiques des sciences de l'homme.

La plupart des utilisations de l'ordinateur dans les humanités numériques qui vont au-delà du stockage et de la recherche d'informations se limitent encore au comptage, aux fréquences et aux statistiques simples. L'interprétation des diagrammes ainsi produits se fait alors le plus souvent sur le tas. Mais ce qui est peut-être encore plus important d'un point de vue épistémologique, et que l'on oublie généralement de mentionner, c'est que les entités à compter, ou le « découpage du phénomène », ne sont généralement définies que par la tradition ou l'intuition :

> C'est que le découpage des faits humains présente une difficulté spécifique. Les phénomènes ont ici un *sens* immédiat, ce qui veut dire qu'ils font spontanément partie d'un univers d'actions valorisées et orientées, soit dans la conscience d'un individu, soit dans l'organisation et le fonctionnement d'une collectivité qui se donne comme un tout, alors même que les liaisons de ce tout nous échappent. Ce sens est véhiculé par le langage pour le sujet parlant de chaque groupe social, et c'est lui qui constitue pour nos consciences d'acteurs l'essence même du fait humain donné. (Granger, 1967, p. 64)

Comme mentionné ci-dessus, il existe des applications plus avancées des mathématiques non numériques dans les humanités numériques, en particulier dans le domaine de l'analyse des réseaux, qui s'appuient sur la théorie des graphes. Cependant, il semble que ces applications soient principalement des emprunts fragmentaires à d'autres disciplines, plutôt que les signes d'une volonté d'embrasser véritablement les mathématiques en tant que science des *patterns*. Pour reprendre un point de l'introduction, je pense que le refus de définir rigoureusement la portée et l'ambition des humanités numériques est l'un des obstacles les plus importants à l'adoption des mathématiques considérées sous cet angle steenéen, si ce n'est le plus important. Cependant, si nous comprenons les humanités numériques comme la construction de modèles formels dans les sciences humaines, il devient évident que nous devons chercher des formalismes appropriés.

L'histoire des sciences humaines de Rens Bod (2013) n'est, comme toujours, qu'un récit possible; quoi qu'il en soit, il ne s'agit pas de

nier que ce que nous appelons aujourd'hui les sciences humaines a toujours été motivé par la recherche de « patterns and principles ». Si nous voulons formaliser cette recherche, ce qui est selon moi la raison d'être des humanités numériques, il est évident que nous devons nous tourner vers la science des *patterns* pour obtenir de l'aide.

5 Enjeux et orientations futures

Il ne fait guère de doute que les idées exposées ci-dessus sont rejetées par la majorité des chercheurs en sciences humaines, et même le courant dominant des humanités numériques ne se lasse pas de souligner que les conflits entre le calcul et l'épistémologie traditionnelle des disciplines des sciences humaines sont non seulement toujours irrésolus, mais en fait « fundamentally unresolvable » (Dobson, 2019, p. 6). L'opposition des sciences humaines à une supposée « réduction des phénomènes aux calculs » que Granger a observée comme un préjugé qui entachait l'appréciation de la formalisation dans les sciences humaines, est encore largement diffusée aujourd'hui. De plus, l'utilisation des ordinateurs tant pour l'automatisation que pour des approches plus avancées nécessite une approche « structuraliste » au sens d'Abraham Moles (1995, p. 141), qui expliquait que « l'hypothèse structurale est basée sur l'idée qu'il est *toujours* possible, et souvent utile, de considérer la réalité comme la combinaison d'éléments ou d'‹ atomes › reconnaissables appartenant à un faible nombre de types et combinés selon un certain nombre, connaissable, de lois ou règles qui constituent précisément ce qu'on appelle la structure ».

Il serait déraisonnable de s'attendre à ce que la « mathématique originale » apparaisse soudainement complètement formée. Resnik (1999, p. 258) souligne que les normes de formalisabilité des mathématiques contemporaines sont trop élevées pour être respectées par la science contemporaine : « Many areas of science make essential use of counterfactual conditionals, causal statements, and vague predicates, none of which can be formalized within contemporary mathematics. ». Ceci est particulièrement vrai pour les sciences humaines et sociales, et mérite d'être souligné à nouveau, parce que les problèmes de recherche dans le monde humain sont intrinsèquement (et ontologiquement) complèxes.

En tant que tels, ces problèmes devraient également constituer un défi pour les mathématiciens. Il faudra voir s'il y a, comme le pensait Kemeny (1959, p. 579), « toutes les raisons de s'attendre à ce que les diverses sciences sociales servent d'incitation au développement de nouvelles grandes branches des mathématiques »; une telle entreprise nécessiterait certainement des interlocuteurs issus des humanités numériques. Sa prédiction selon laquelle « un jour, le chercheur en sciences sociales théoriques devra connaître plus de mathématiques que le physicien n'a besoin d'en connaître aujourd'hui » semble improbable à première vue, mais si nous sommes capables de progresser dans la formalisation, par exemple, de l'incertitude telle qu'elle est rencontrée dans les sciences humaines (voir p. ex. Piotrowski, 2023), il est probable qu'elle nécessitera beaucoup de mathématiques.

D'un autre côté, on pourrait être d'avis qu'avec l'émergence de l'IA générative (principalement les grands modèles de langage), toute cette abstraction et cette formalisation seront inutiles, puisque les ordinateurs on désormais accès à ce que Granger (1967, p. 64) appelle le « sens immédiat » qui est « véhiculé par le langage [...] et c'est lui qui constitue pour nos consciences d'acteurs l'essence même du fait humain donné. » Quoi qu'il en soit, Moles (1995, p. 287) a certainement raison en affirmant que « [s]avoir penser *avec* l'ordinateur [...] est bien une nouvelle situation de l'esprit que n'ont connue ni Leibniz, ni Descartes, Hume ou Locke. » Ce que nous ne savons pas, en revanche, c'est si l'ordinateur continuera à être le « gardien de la vérité logique ».

6 Conclusion

Le point de départ de cet essai est l'affirmation de Granger d'endosser la tâche consistant à montrer la pensée formelle à l'œuvre dans les sciences humaines, « non pas seulement comme réduction des phénomènes aux calculs, mais aussi comme invention de structures nouvelles, voire même d'une mathématique originale » (Granger, 1967, p. 19).

Granger nous donne un indice en se référant à Kemeny (1959), et si nous adoptons un point de vue historique, nous pouvons effectivement affirmer que les mathématiques ont évolué de l'art de la mesure à une

(ou même *la*) science des *patterns* (Steen, 1988). Cette conception des mathématiques les rend beaucoup plus pertinentes pour les sciences humaines et sociales, qui sont, comme le dit Bod (2013), à la recherche de « modèles et de principes ». En ce sens, la « mathématique originale » de Granger serait donc bien le fondement des humanités numériques qui s'intéressent à la construction de modèles formels dans les sciences humaines.

Comme l'a fait remarquer Hamming (1962, p. vii), « The purpose of computing is insight, not numbers », et Thom (1991) nous rappelle que « prédire n'est pas expliquer ». Ainsi, si l'utilisation d'ordinateurs pour automatiser le comptage, le calcul et la production de statistiques n'est pas « erronée », elle n'exploite les possibilités que dans une faible mesure. Comme l'a noté Borillo (1985, p. 5), relier l'informatique à son socle conceptuel (logico-mathématique) « fera évoluer également les cadres théoriques et les référents épistémologiques des sciences de l'homme. »

Pour avancer sur cette voie, il faut des chercheurs véritablement pluridisciplinaires comme François Bavaud.

Références

Bavaud, F. & Métrailler, C. (2023). A (dis)similarity index for comparing two character networks Based on the Same Story. In Rochat, Y., Métrailler, C., & Piotrowski, M. (éd.), *Proceedings of the Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022)*, pages 33–42. CEUR Workshop Proceedings.

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press, Oxford.

Borillo, M. (1985). *Informatique pour les sciences de l'homme: Limites de la formalisation du raisonnement*. Mardaga, Bruxelles.

Comte, A. (1892). *Cours de philosophie positive*, volume 1. Société positiviste, Paris, 5ème édition.

Couturat, L. (1904). Les principes des mathématiques. *Revue de Métaphysique et de Morale*, 12(1):19–50.

Dobson, J. E. (2019). *Critical digital humanities: The search for a methodology*. University of Illinois Press, Champaign, IL.

Egloff, M. & Bavaud, F. (2018). Taking into account semantic similarities in correspondence Analysis. In Piotrowski, M. (éd.), *Proceedings of the Workshop on Computational Methods in the Humanities (COMHUM 2018)*, pages 45–51. CEUR Workshop Proceedings.

Granger, G.-G. (1967). *Pensée formelle et sciences de l'homme*. Aubier-Montaigne, Paris. Nouvelle éd. augmentée d'une préface.

Hamming, R. W. (1962). *Numerical methods for scientists and engineers*. McGraw-Hill, New York, NY.

Kemeny, J. G. (1959). Mathematics without Numbers. *Daedalus*, 88(4):577–591.

Moles, A. A. (1995). *Les sciences de l'imprécis*. Seuil, Paris.

Piotrowski, M. (2023). Uncertainty as Unavoidable Good. Center for Uncertainty Studies Working Papers 5, Universität Bielefeld, Center for Uncertainty Studies (CeUS), Bielefeld.

Piotrowski, M. & Xanthos, A. (2020). Décomposer les humanités numériques. *Humanités numériques*, 1.

Resnik, M. D. (1999). *Mathematics as a science of patterns*. Oxford University Press, Oxford.

Speiser, A. (1952). *Die mathematische Denkweise*. Birkhäuser, Basel, 3ème édition.

Steen, L. A. (1988). The Science of Patterns. *Science*, 240(4852):611–616.

Thom, R. (1991). *Prédire n'est pas expliquer: Entretiens avec Émile Noël*. Flammarion, Paris, 2ème édition.

Wirth, N. (1976). *Algorithms + data structures = programs*. Prentice-Hall, Englewood Cliffs, NJ.

Cahiers du CLSL, n° 69, 2025, pp. 219-237

« May you find peace on your journey » : une approche comparative des mondes de FromSoftware

Loris Rimaz & Coline Métrailler
Université de Lausanne
{loris.rimaz,coline.metrailler}@unil.ch

Résumé

FromSoftware, un studio de développement japonais mondialement connu pour sa série de jeux dits « *SoulsBorne* », propose des mondes à l'ambiance et au ton constants à travers les opus, mais dont l'expérience spatiale des joueuses et des joueurs varie considérablement d'un jeu à l'autre. Afin de mettre en lumière ces différences, nous proposons une représentation en graphes, ce qui permet non seulement de souligner les différences intuitives entre les univers de FromSoftware, mais également d'interroger les conséquences et les effets que ces différences structurelles induisent. À travers l'exemple concret des jeux FromSoftware, ce papier présente ainsi une méthode d'analyse formelle des mondes vidéoludiques.

1 Introduction

Il est aisé de voir les similarités entre les jeux du studio de développement japonais FromSoftware. Depuis 2009, l'équipe raffine le design de ses jeux de rôle d'action (A-RPG) ; la publication de *Demon's Souls* marque le début d'une série de jeux dits « *SoulsBorne* » qui partagent de nombreux points communs. La difficulté brutale, parfois injuste, des séquences de jeu en a fait une saga renommée : véritable rituel initiatique, la complétion d'un A-RPG de FromSoftware devient un symbole d'appartenance à une classe de joueurs et joueuses persévérantes et aguerries. Pourtant, les similarités entre les jeux ne se limitent pas à leur *gameplay*. Les histoires qu'ils racontent partagent une narration opaque dont seuls quelques éléments sont dévoilés avant de donner le

contrôle aux joueurs et joueuses, et il est facile de passer à côté des récits complémentaires à débloquer au fil de la partie.

Si la genèse des univers dans lesquels les jeux se déroulent n'échappe pas à cette hermétisme, les mondes qu'ils contiennent partagent des caractéristiques formelles évidentes : ils sont en crise, abandonnés par leurs Dieux ou leurs Seigneurs, à la fin de leur ère. La désolation et l'abandon sont palpables, et le conflit et la violence sont les seuls modes d'interaction restants avec le peu de vie qui s'y trouve encore. Cependant, l'agencement spatial des mondes de FromSoftware change considérablement d'un jeu à l'autre. Par exemple, le Royaume de Drangleic (*Dark Souls II*) semble plus linéaire, plus étroit, que le Royaume de Lordran (*Dark Souls*). Les deux s'organisent pourtant autour d'un noyau central à partir duquel les joueurs et joueuses explorent les environs, mais les régions de Drangleic paraissent plus discontinues et moins interconnectées que les régions de Lordran – à certains égards, Lordran pourrait être considéré comme plus cohérent. De plus, la question de la liberté qu'offrent les mondes de jeu est au cœur de nombreux débats entre les fans de la saga. En particulier, le sentiment qui semble se démarquer est que *Dark Souls II* et *Dark Souls III* seraient particulièrement linéaires (Binkt, 2016; DarkStar643, 2016; deleted user, 2015; saito200, 2017; SCP, 2019; SJIS0122, 2023), là où *Dark Souls* et *Bloodborne* offrent des mondes fortement interconnectés (AdPuzzleheaded8723, 2024; SarcasticLizard, 2016).

Ces différences intuitives dans la structure de l'espace amènent à une compréhension et une expérience spatiale différentes. Afin de les mettre en lumière, nous passons par une représentation en graphes orientés des mondes de FromSoftware. Cette étape d'abstraction permet non seulement de rendre compte de la structure de leur agencement spatial, mais également de mettre en lumière la progression dans, et ainsi l'expérience de, ces espaces virtuels. Notre objectif avec cette méthode d'analyse formelle de mondes vidéoludiques est d'estimer, dans un premier temps, si les différences intuitives entre les univers de FromSoftware sont effectivement le produit d'un agencement spatial différent. Par la suite, nous évaluons la dimension diachronique, c'est-à-dire l'évolution, de ces mondes, et soulignons les logiques de *design* qui ont motivé certains

changements. Enfin, nous interrogeons les conséquences et les effets que ces différences structurelles induisent : en particulier, nous explorons la disparité entre l'espace tel qu'il est vécu et l'espace tel qu'il est réellement agencé. La passion notoire de François Bavaud pour les données spatiales et les analyses de graphes nous a inspiré cette recherche, et nous espérons que nos résultats sauront éveiller son intérêt. Les données nécessaires à ces différentes analyses, présentées à la section suivante, ont été nettoyées à la main et sont désormais accessibles sur un répertoire dédié[1] pour de futures recherches.

2 Méthodologie

Dans le périmètre de ce projet, nous nous concentrons sur les opus suivants de FromSoftware :

— *Dark Souls* (2011)
— *Dark Souls II* (2014)
— *Bloodborne* (2015)
— *Dark Souls III* (2016)
— *Elden Ring* (2022)

La série des *Dark Souls* est particulièrement propice pour l'analyse de l'évolution du *design* des mondes de FromSoftware. *Bloodborne* et *Elden Ring* sont utiles à des fins de comparaison et de nuance. L'omission de *Demon's Souls* (2009) et *Sekiro : Shadows Die Twice* (2019) sera justifiée dans les paragraphes qui suivent.

Un graphe orienté est composé de sommets et d'arcs qui relient ces sommets entre eux dans une direction déterminée. Afin de représenter les mondes de FromSoftware sous la forme de graphes orientés, il faut déterminer les points du terrain qui feront office de sommets ; les arcs symboliseront ainsi la possibilité de déplacement entre ces emplacements. Les feux de camp, introduits dans *Dark Souls*, émergent rapidement comme les candidats idéaux. Ces éléments, servant de points de contrôle, deviennent une des mécaniques ludiques cruciales des A-RPG de FromSoftware à partir de 2011 : se reposer à un feu de camp

1 https://github.com/DigitalDW/FROMSOFT_NETWORK.

permet de remplir les jauges de vie, endurance (*stamina*) et magie, ainsi que de reconstituer son stock de consommables rechargeables (comme les flasques d'Estus qui régénèrent la vie lorsqu'on les boit). Cette régénération du personnage se fait au coût de la réinitialisation de la position des ennemis en déplacement et à la réapparition de tous ceux qui ont été vaincus depuis le dernier repos (à l'exception des *boss*). Les feux de camp, ainsi que leur équivalent dans *Bloodborne* (lampes) et *Elden Ring* (sites de grâce), sont stratégiquement placés directement avant ou après une zone à explorer. Le choix des feux de camp comme sommets pour les graphes n'est dès lors pas anodin : ils représentent des emplacements-clefs situés aux seuils des zones de chaque jeu[2].

2.1 Données

Les données des sommets ont été collectées sur les Wikis des différents jeux[3] : pour chacun d'entre eux, il est possible de trouver, généralement sous la forme d'un tableau, une liste des points de contrôle qui renseigne leur nom ainsi que la région dans laquelle ils se situent. Les données ont d'abord été extraites manuellement en copiant le `HTML` de la page web dans un fichier `.txt`. Les informations inutiles comme les balises `HTML` ont été supprimées à l'aide d'expressions régulières. Au terme de ces opérations, les données nettoyées ont été formatées en `XML` avant d'être converties en `JSON`. Enfin, un numéro d'identification a été ajouté à chaque entrée à l'aide d'un script en Python. La forme finale d'une entrée dans le fichier `JSON` obtenu est la suivante :

2 *Demon's Souls* ne possède pas d'éléments équivalents, ce qui explique pourquoi nous ne l'avons pas traité.

3 Liste des wikis utilisés :
- *Dark Souls* : `http://darksouls.wikidot.com/bonfires`
- *Dark Souls II* : `http://darksouls2.wikidot.com/bonfires`
- *Bloodborne* : `http://bloodborne.wikidot.com/lamps`
- *Dark Souls III* : `https://darksouls3.wiki.fextralife.com/Bonfires`
- *Elden Ring* : `https://eldenring.wiki.fextralife.com/Sites+of+Grace`

```
{
    "zone": "Cemetery of Ash",
    "name": "Firelink Shrine",
    "id": 1
}
```

Les données des arcs ont été produites à la main en se basant à la fois sur les Wikis des différents jeux et sur notre propre expérience des mondes de FromSoftware[4]. Pour accélérer le processus, un script court en Python permettant de lister les sommets à connecter à un autre a été développé. Seul *Elden Ring* n'a pas pu bénéficier de ce script, ses 313 points de contrôle étant trop nombreux pour s'afficher agréablement dans une liste.

La création des arcs a posé quelques questions de définition : certaines régions représentées par les sommets ne sont pas accessibles dès le début du jeu et ne sont débloquées qu'après avoir accompli certains objectifs ; à l'inverse, la téléportation permet de rejoindre instantanément certaines (ou toutes les) régions déjà visitées. Une définition trop restrictive ou trop permissive des arcs produirait ainsi des graphes déconnectés (si certaines zones sont accessibles plus tard) ou complets (si toutes les zones sont reliables par téléportation), ce qui limiterait l'intérêt des études desdits graphes.

Il a donc été décidé d'ignorer la téléportation, à part dans les rares exceptions où il s'agit de la seule manière de rejoindre une région (on considère alors qu'il s'agit d'un chemin direct vers cette destination). Ainsi, dans le fichier final, un arc entre un sommet d'origine et un sommet de destination n'est présent que si : (1) le sommet de destination est accessible (dès le premier passage ou plus tard) par le sommet d'origine et (2) il représente le plus court chemin entre les deux sommets.

2.2 Mesures

Dans l'optique d'observer une évolution des expériences spatiales entre les différents jeux, mais aussi d'analyser en détail l'importance de

4 Le travail à fournir pour générer les arêtes de *Sekiro : Shadows Die Twice* était ainsi trop important par rapport à nos connaissances du jeu.

chaque lieu au sein d'un même jeu, plusieurs mesures mathématiques ont été sollicitées sur les graphes ainsi engendrés. Notons que dans cette partie exploratoire, nous avons pris en compte une version non orientée des graphes (avec l'espoir toutefois que les données seront utiles à de futures recherches plus approfondies), c'est pourquoi nous parlerons d'*arêtes* et non d'*arcs* dans la suite de cet article.

Les mesures qui concernent l'intégralité du graphe sont les suivantes :

— **Diamètre** : distance maximale (en nombre d'arêtes parcourues) à l'intérieur du graphe.
— **Excentricité moyenne** : moyenne des distances maximales entre chaque sommet et le reste du graphe.
— **Densité** : proportion d'arêtes existantes sur les arêtes possibles.
— **Fermeture triadique** : proportion de triangles existants sur tous les triangles possibles.
— **Distance moyenne** : moyenne des distances entre chaque paire de sommets du graphe.
— **Taille maximale de clique** : nombre maximal de sommets interconnectés.

Les valeurs ainsi récoltées pour les différents jeux permettent de décrire l'éloignement relatif entre tous les lieux étudiés (diamètre, distance moyenne, densité, chemin le plus court), ainsi que la possibilité de prendre différents chemins pour accéder au même point d'arrivée (fermeture triadique et taille des cliques).

En outre, l'importance relative des lieux peut se mesurer avec plusieurs approches de centralité. Dans le cadre de cette recherche, nous avons choisi de mettre en perspective trois mesures centrées sur les sommets :

— **Degré** : nombre d'arêtes associées au sommet.
— **Centralité de proximité** : inverse de l'excentricité d'un sommet.

$$C_B(x) := \frac{1}{\sum_y d(x,y)},$$

où x et y sont des sommets, et $d(x,y)$ est la distance entre x et y.
— **Centralité intermédiaire** : proportion de plus courts chemins

qui passent par le sommet.

$$C_I(x) := \frac{\sum \sigma_{st}(x)}{\sum \sigma_{st}},$$

où x, s et t sont des sommets, σ_{st} est le nombre total de plus courts chemins de s à t, $\sigma_{st}(x)$ est le nombre de plus courts chemins entre s et t passant par x et les sommes $\sum$ sont calculées sur tous s, t tels que $s \neq x \neq t$.

La combinaison de ces trois scores permet de déterminer, pour chaque lieu, à quel point il est relié au reste du graphe (degré), à quel point il est relié à d'autres lieux importants (centralité de proximité), et à quel point il sert de relais entre des zones entières de la carte (centralité intermédiaire).

2.3 Génération des résultats

Sur la base des données nettoyées, les graphes ont fait l'objet de deux traitements parallèles.

Avec l'aide de la librairie `networkx`, ils ont été construits en Python pour en calculer toutes les mesures présentées à la section 2.2. Les résultats ont ensuite été reportés dans des tableaux : un tableau général (voir Tab. 1, p. 237) et des tableaux complets des scores individuels de chaque sommet pour chacun des jeux[5], afin d'obtenir des informations détaillées sur tous les lieux étudiés.

Parallèlement, les données ont été traitées en JavaScript, à l'aide de la librairie `D3` (*Data Driven Document*), à des fins de visualisation. Deux modes de visualisation ont été développés : la vue orientée et la vue simplifiée.

La première affiche des arêtes courbées afin de mettre en évidence la dimension orientée du graphe : deux arêtes relient les sommets entre lesquels un aller-retour est possible, tandis qu'une seule arête est présente si l'accès est uni-directionnel.

La deuxième n'affiche qu'une arête par connexion, ne proposant qu'une légère variation dans la couleur (plus foncée) si le lien est bidirectionnel. Cette vue propose d'élargir les arêtes en fonction de la

5 Ces tableaux sont à retrouver sur le répertoire du projet.

distance entre les sommets dans les données. En d'autres termes, si le sommet `1` est connecté au sommet `2`, leur arête sera plus étroite que celle qui connecte les sommets `2` et `33`, p. ex. ; puisque les sommets sont classés selon la progression recommandée dans les jeux par leur Wiki, la largeur des arêtes peut indiquer une connexion qui n'est disponible que plus tard. Cependant, cette mesure n'est jamais exploitée telle quelle dans les résultats et a surtout été utilisée à des fins de correction dans la génération des données.

Enfin, les deux vues proposent un même code couleur pour les sommets : chaque région se voit attribuer une couleur. Ainsi, les sommets d'une même couleur appartiennent à la même région.

3 Résultats et discussion

3.1 Représentations de l'espace

Les espaces virtuels créés pour les jeux vidéo sont des espaces construits pour déployer une pratique ludique selon les modes d'interactions définis par leurs auteurs et autrices. Le philosophe Henri Lefebvre propose un modèle tripartite pour la production de l'espace social : la pratique spatiale, les représentations de l'espace et les espaces de représentations – le perçu, le conçu et le vécu. Si « la pratique spatiale d'une société se découvre en déchiffrant son espace » (Lefebvre, 1981, p. 48), alors l'espace d'un jeu vidéo est *a priori* paradoxal. En effet, pour Mathieu Triclot (2012), « s'il y a réellement espace dans les jeux vidéo, il faut donc dire qu'en lui la représentation précède la pratique, l'artifice la nature, l'abstrait le concret. » (p. 223). En d'autres termes, l'espace vidéoludique n'est pas issu d'une pratique, mais plutôt d'une conception. Triclot affirme même que « le processus vivant de production de l'espace que décrit Lefebvre [...] semble totalement étranger au jeu vidéo » (2012, p. 223). Cette affirmation semble aller à l'encontre des processus de développement de jeux vidéo tels qu'ils sont décrits dans leurs post-mortems[6] (Romero & Hall, 2016; Yu & Hull, 2017) : quand bien même l'espace des jeux vidéo correspond plutôt à l'espace conçu – la

6 Présentations ou articles dans lesquels des auteurs et autrices de jeux reviennent sur le processus de développement, expliquant notamment leurs choix de *design* et de conception.

représentation d'un espace –, la pratique attendue est testée pendant le processus de développement. Ainsi, la pratique spatiale, et donc la pratique attendue, sont bel et bien évidentes en déchiffrant l'espace, précisément parce qu'elles y ont été déployées à de nombreuses reprises durant la création du jeu.

Par exemple, chaque jeu possède au moins un lieu perçu comme central. Il s'agit d'un endroit où les joueurs et les joueuses peuvent, en toute sécurité, améliorer leur personnage et leur équipement, ou interagir avec des personnages non-joueurs (PNJ). Ces emplacements, souvent référencés par le terme anglais *hub*, sont des centres où sont concentrées d'autres mécaniques de jeu que celles déployées durant les phases d'exploration ou de combat. Le *design* de ces lieux[7] indique qu'ils sont sûrs et propices à une pratique différente.

Selon la définition tripartite de Lefebvre, les graphes sont des représentations de l'espace des mondes de FromSoftware. Ainsi, sans surprise, ils diffèrent considérablement de l'agencement spatial des mondes tels qu'ils sont dans les jeux. Le fait qu'un graphe ne représente pas la verticalité présente dans un espace en trois dimensions explique en partie cette différence. Cet aspect est particulièrement remarquable avec *Dark Souls III* : bien que la plupart des régions se superposent physiquement, le fait qu'elles ne soient pas interconnectées leur donne un aspect complètement linéaire dans notre représentation. De la même façon, les longues branches du graphe de *Dark Souls II* ne transmettent aucune information sur le dénivelé ou l'éventuelle superposition des régions. L'intérêt principal des graphes est de confronter l'espace perçu avec l'espace conçu. En d'autres termes, nos représentations de l'espace ne reflètent pas l'agencement de l'espace mais plutôt une forme de progression dans les mondes de FromSoftware.

3.2 Évolution des mondes

3.2.1 La série *Dark Souls*

Les trois jeux *Dark Souls* forment un triptyque intéressant puisqu'ils appartiennent au même univers et, en se basant simplement sur les

7 Le *design* peut inclure ici des aspects visuels (architecture, disposition des éléments, voire même la météo), sonores (musique différente), ainsi que la présence de PNJ amicaux.

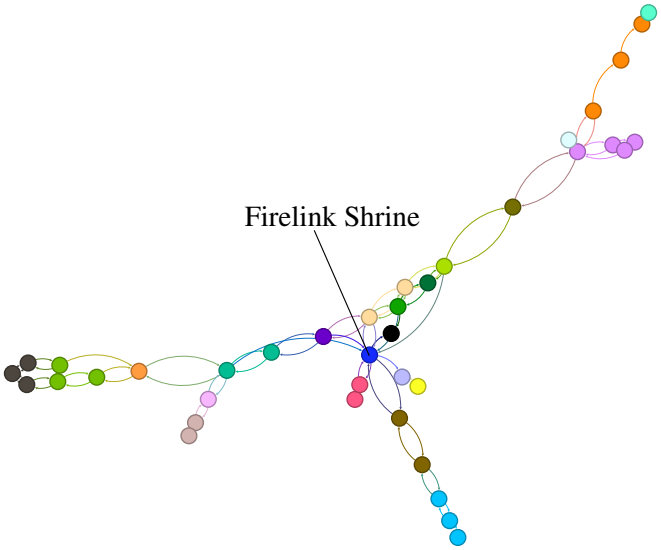


FIGURE 1 – Graphe orienté pour *Dark Souls*.

graphes, il semble clair que leur évolution correspond aux ressentis de la communauté de joueurs et de joueuses quant à leur structure (Ad-Puzzleheaded8723, 2024; DarkStar643, 2016; SarcasticLizard, 2016; SJIS0122, 2023). Comme le montre la figure 1, *Dark Souls* propose un groupe de sommets en son centre, duquel partent trois branches principales, qui amènent elles-mêmes vers d'autres ensembles de sommets. Le sommet central tel qu'il est vécu dans le jeu est « Firelink Shrine », et il se situe au centre du groupe principal. Au contraire, *Dark Souls II* (Fig. 2) semble, au premier abord, proposer un disposition symétrique avec deux moitiés à trois branches chacune. Cependant, sachant que le sommet qui est l'équivalent de « Firelink Shrine » de *Dark Souls* est « The Far Fire », le graphe montre plutôt cinq chemins qui quittent le nœud central. En d'autres termes, le graphe représente en réalité une disposition en étoile. Enfin, *Dark Souls III* s'organise autour d'une artère centrale à partir de laquelle plusieurs branches, relativement courtes, se séparent (Fig. 3). À y regarder de plus près, il est surprenant de constater que « Firelink Shrine », le sommet vécu comme central par les

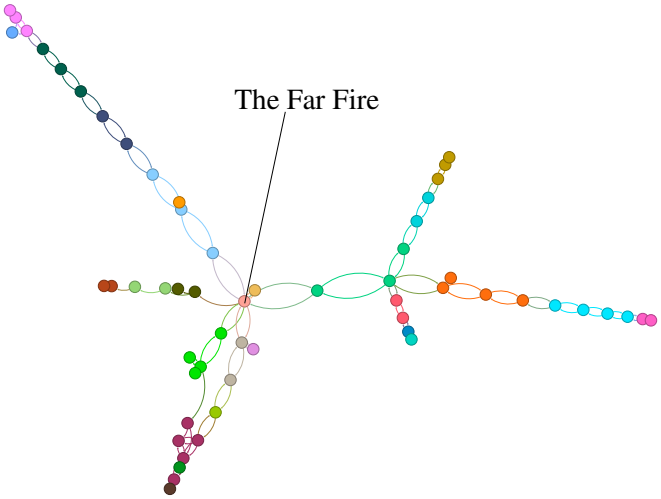


FIGURE 2 – Graphe orienté pour *Dark Souls II*.

joueurs et les joueuses, se situe près d'une des extrémités du graphe, sur une branche. L'évolution est claire : les jeux tendent vers une forme de linéarité, dans un premier temps organisée autour d'un sommet central (*Dark Souls II*) avant de s'étendre à l'ensemble de la progression (*Dark Souls III*).

Ce constat est renforcé par les mesures. Si l'on s'intéresse au diamètre des trois graphes des *Dark Souls*, on peut voir qu'il ne fait qu'augmenter, confirmant cette impression d'une linéarité grandissante (le nombre de sommets étant identique entre *Dark Souls II* et *Dark Souls III*). L'excentricité moyenne et la distance moyenne augmentent elles aussi, et la densité est extrêmement faible dans les trois cas (avec respectivement 6.6%, 3.8% et 4.2% d'arêtes existantes). Quant à la question des cliques, les mesures révèlent une taille maximale de 3 pour chacun des jeux : s'il est ainsi possible de faire des triangles dans certaines zones du graphe, il n'existe aucun groupe plus grand, ce qui laisse également penser que les chemins ont été prévus pour être parcourus dans un ordre précis et sans grande liberté dans l'itinéraire.

En regardant de plus près les mesures de chaque sommet, on peut en outre constater que si « Firelink Shrine » (*Dark Souls*) et « The Far Fire »

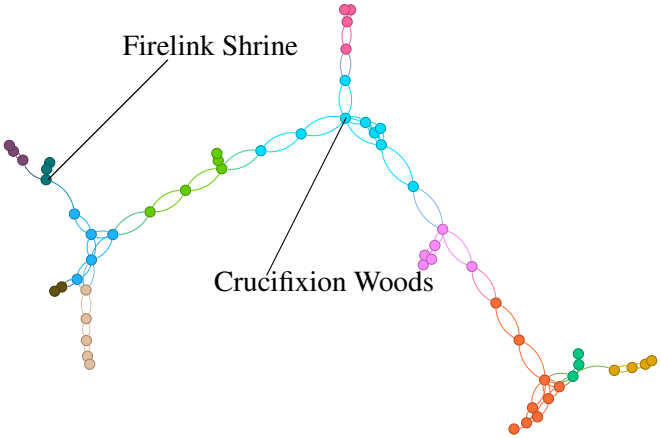


FIGURE 3 – Graphe orienté pour *Dark Souls III*.

(*Dark Souls II*) ont des valeurs cohérentes avec leur position de « centre d'étoile » (comme mentionné plus haut) avec un degré, une centralité de proximité et une centralité intermédiaire élevés, la situation est moins claire pour *Dark Souls III* avec plusieurs sommets qui se partagent les valeurs maximales pour ces trois mesures. En particulier, la centralité intermédiaire nous révèle que si dans les deux premiers jeux, seuls 3 sommets dépassent une centralité intermédiaire de 0.40, on retrouve 11 sommets avec un tel score dans *Dark Souls III*, soit 19% de la totalité des sommets du graphe, ce qui marque bien la forte linéarité de cet opus (beaucoup de sommets étant des passages obligatoires pour un grand nombre de chemins).

Cette évolution pourrait s'expliquer par des changements dans la direction des équipes de développement. *Dark Souls* a été dirigé par Hidetaka Miyazaki, tandis que pour *Dark Souls II*, Tomohiro Shibuya a d'abord géré le projet avant de passer le flambeau à Yui Tanimura. Ceci expliquerait en partie les points communs et différences structurelles entre les Royaumes de Drangleic et Lordran : l'organisation autour d'un point central au cœur du monde est un héritage direct de *Dark Souls* et pourrait traduire une volonté de s'inscrire dans la lignée du premier jeu, tandis que la progression beaucoup plus linéaire sur les branches serait

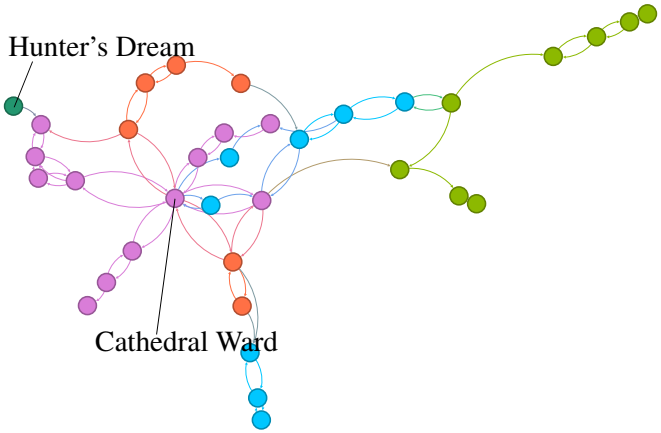


FIGURE 4 – Graphe orienté pour *Bloodborne*.

issue de la nouvelle direction de l'équipe. *Dark Souls III* est le produit d'une co-direction de Miyazaki et Tanimura. Ainsi, la linéarité pourrait s'expliquer par l'influence de Tanimura dans le processus de conception de l'espace. De plus, le système de téléportation entre tous les feux de camp, repris directement de *Dark Souls II*, expliquerait la disparition de « Firelink Shrine » dans les mesures de centralité : il n'était plus strictement nécessaire de le placer au centre du monde, puisqu'on peut désormais s'y rendre depuis n'importe quel feu de camp[8].

3.2.2 *Bloodborne*

Comme nous allons le voir, la représentation de l'espace en graphe orienté de *Bloodborne* (Fig. 4) met en lumière, d'une part, un retour à la structure spatiale de *Dark Souls* et, d'autre part, la mise à l'écart du *hub* vécu comme central que l'on retrouve dans *Dark Souls III*.

En effet, son graphe partage plus de similarités avec celui de *Dark Souls* (Fig. 1) qu'avec celui de *Dark Souls II* (Fig. 2) : il possède notamment un groupe central plus interconnecté et plusieurs branches

8 Cette dynamique est probablement directement inspirée de *Bloodborne*, paru un an avant *Dark Souls III*. En effet, Fromsoftware s'était déjà libéré de la contrainte de placer au centre du monde le sommet (« Hunter's Dream » dans *Bloodborne*) vers lequel les joueurs et les joueuses reviennent pour améliorer leur personnage et équipement.

linéaires qui s'en détachent. De plus, les mesures indiquent que *Bloodborne* et *Dark Souls* sont presque deux fois plus denses et interconnectés que *Dark Souls II* [9]. Son sommet central, avec un degré presque autant élevé que « Firelink Shrine » (*Dark Souls*), est « Cathedral Ward » [10].

Ainsi, à l'instar de *Dark Souls III*, le sommet qui ressort comme central n'est pas le sommet vécu comme un *hub* du jeu, ce qui illustre bien l'intérêt principal de nos représentations de l'espace, à savoir la possibilité de comparer l'espace à sa réception. En effet, le sommet vécu comme un *hub* par les joueurs et les joueuses est « Hunter's Dream » (jancasellass, 2021). Après tout, il s'agit du seul endroit où l'on peut dépenser les points accumulés pour augmenter les caractéristiques de son personnage, améliorer son équipement ou changer certains attributs. De plus, chaque lanterne (point de contrôle) rencontrée dans le monde du jeu ne permet que de se reposer ou de se téléporter vers « Hunter's Dream ». Ainsi, « Hunter's Dream » serait attendu comme sommet central idéal, puisqu'il connecte tous les sommets entre eux. Cependant, il est physiquement déconnecté de Yharnam et de ses environs. De fait, notre représentation de l'espace fait ressortir « Cathedral Ward » comme étant le sommet central : il possède le degré le plus élevé, en plus d'avoir les plus hautes valeurs pour les deux mesures de centralité.

Sur un plan physique, c'est bien « Cathedral Ward » qui est au cœur de Yharnam et qui connecte les différentes régions du monde[11]. Espace de répit au sein de Yharnam – plusieurs PNJ s'y réfugient au fil du jeu – « Cathedral Ward » fait figure de *hub* narratif, là où « Hunter's Dream » est un *hub* mécanique. Ce que la différence entre la perception de « Hunter's Dream » et sa place dans notre conception de l'espace de *Bloodborne* souligne avant tout, c'est la subjectivité de l'expérience. Un

9 *Bloodborne* : densité de 7.2% et plus court chemin moyen de 4.104 ; *Dark Souls* : densité de 6.6% et plus court chemin moyen de 4.907 ; *Dark Souls II* : densité de 3.8% et plus court chemin moyen de 7.815.

10 Il possède également les plus hautes valeurs de centralité (de proximité et intermédiaire).

11 D'ailleurs, si tous les liens qui représentent des déplacements atypiques comme le carrosse qui transporte le personnage de « Hemwick Channel Lane » / « Witch's Abode » à « Forsaken Castle Cainhurst » ou la téléportation obligatoire entre « Lecture Building 2nd Floor » et « Nightmare of Mensis » étaient retiré du graphe, la centralité de « Cathedral Ward » ne serait que renforcée.

joueur ou une joueuse qui attribue plus d'importance à la mécanique de jeu attachera plus d'importance symbolique aux environnements de « Hunter's Dream », là où celui ou celle qui a un attrait particulier pour (la spatialisation de) la narration vivra « Cathedral Ward » comme le cœur du jeu.

3.2.3 *Elden Ring*

Elden Ring (Fig. 5) représente une évolution des mondes relativement linéaires et fermés des précédents jeux de FromSoftware vers un monde ouvert. La représentation en graphe d'un tel jeu est beaucoup plus complexe à mettre en place, notamment à cause de l'augmentation massive du nombre de sommets à connecter (entre trois et cinq fois plus d'arêtes à produire par sommet). Le graphe obtenu avec notre méthode est déjà très riche mais probablement trop léger pour en tirer de grandes conclusions. Il est cependant intéressant de souligner qu'en l'état, c'est un graphe non connexe : en effet, dans la tradition de *Bloodborne* et *Dark Souls III*, il y a un sommet vécu comme central (« Table of Lost Grace ») qui n'est accessible qu'en se téléportant. La différence avec les jeux précédents est qu'il n'y a pas un moment précis où les joueurs et les joueuses y accèdent pour la première fois. Les conditions pour y arriver varient, et nous avons estimé qu'il était impossible de déterminer à quel sommet le rattacher sans un accès à des données télémétriques des parcours des joueurs et des joueuses.

Une autre observation intéressante est la linéarité du dernier tiers du jeu. En effet, les deux premiers tiers forment un cercle où nous retrouvons plusieurs regroupements de sommets, tandis que le dernier tiers est représenté par une branche qui se sépare du cercle (en haut à gauche de la figure 5). Cette branche possède un embranchement majeur (qui est d'ailleurs optionnel) ainsi que quelques petites branches annexes et présente une forte séquentialité par rapport au reste du graphe. Une majorité des sommets s'y trouvant possèdent un faible degré (rarement plus que 1) ainsi que des valeurs de centralité basses[12].

12 Il y a seulement quelques sommets dont la centralité intermédiaire est supérieure à 0.1, et la valeur ne fait que diminuer plus le sommet est éloigné du cercle.

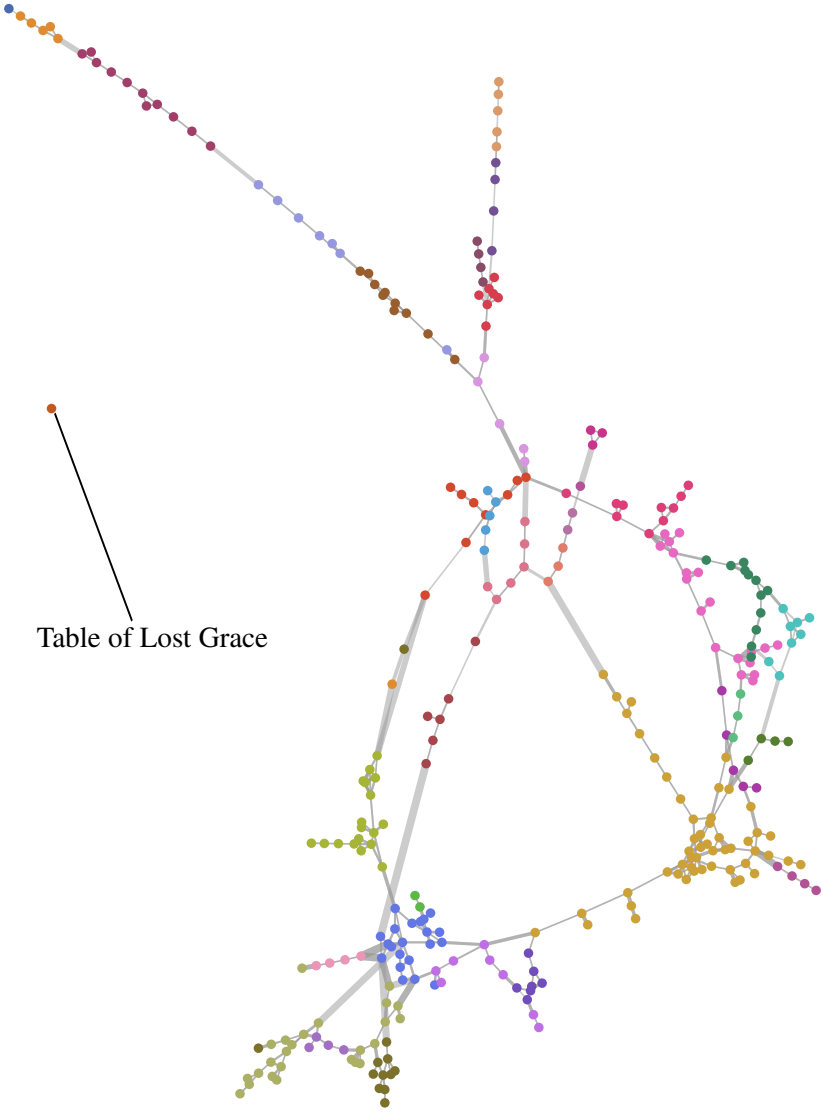


FIGURE 5 – Graphe en vue simplifiée pour *Elden Ring*.

4 Conclusion

Les mondes créés par FromSoftware ont évolué au fil des jeux. La représentation de l'espace en graphes orientés confirme le ressenti des joueurs et des joueuses : *Dark Souls* est moins linéaire que *Dark Souls II*, lui-même moins linéaire que *Dark Souls III*. Cependant, les graphes ne traduisent que la progression dans le monde et non la progression au sein des zones, qui peut elle-même varier de l'une à l'autre et entre les jeux. La méthode choisie met également en lumière certaines limitations : la richesse d'*Elden Ring* n'est pas nécessairement apparente, si ce n'est par le nombre de sommets dans le graphe. Ceci étant dit, ces représentations de l'espace permettent de confronter l'agencement spatial des mondes de FromSoftware – le perçu – et la signification que les joueurs et les joueuses leur attribuent – le vécu – avec la progression qu'ils rendent possible – le conçu. Il serait dès lors intéressant de modéliser l'espace de *Sekiro : Shadows Die Twice* afin d'analyser ses emprunts et son apport à la conception de l'espace dans les A-RPG de FromSoftware. Enfin, puisque les données sont rendues accessibles, il serait utile de les enrichir afin de proposer d'autres analyses : les graphes pourraient être rendus dynamiques, on pourrait également déterminer et analyser les plus courts chemins pour traverser les jeux, et des données télémétriques permettraient d'analyser les flux des joueurs et des joueuses entre les sommets (et ainsi construire des chaînes de Markov, entre autres).

Références

AdPuzzleheaded8723 (2024). Which Soulsborne game has the best interconnectivity ? https://www.reddit.com/r/fromsoftware/comments/1966u86/which_soulsborne_game_has_the_best/. [Post Reddit].

Binkt (2016). Why does this game feel so linear/straight ? https://gamefaqs.gamespot.com/boards/168567-dark-souls-iii/74078329. [Post Reddit].

DarkStar643 (2016). Is Dark Souls 3 more linear than Dark Souls 2 ? https://gamefaqs.gamespot.com/boards/168566-dark-souls-iii/73670990. [Post Reddit].

deleted user (2015). Do you think Dark souls 2 is more linear than 1 ? https://www.reddit.com/r/darksouls/comments/3ymxhp/do_you_think_dark_souls_2_is_more_linear_than_1/. [Post Reddit].

FromSoftware (2011). Dark Souls. Namco Bandai Games.

FromSoftware (2014). Dark Souls II. Namco Bandai Games.

FromSoftware (2015). BloodBorne. Sony Interactive Entertainment.

FromSoftware (2016). Dark Souls III. Bandai Namco Entertainment.

FromSoftware (2022). Elden Ring. Bandai Namco Entertainment.

jancasellass (2021). Best hub area in your opinion ? https://www.reddit.com/r/Eldenring/comments/omqt4q/best_hub_area_in_your_opinion/. [Post Reddit].

Lefebvre, H. ([1974] 1981). *La production de l'espace*. Anthropos, Paris, 2ème édition.

Romero, J. & Hall, T. (2016). "Doom :" a classic game postmortem. https://www.youtube.com/watch?v=NnkCujnYNSo. [Vidéo].

saito200 (2017). Is this game too linear ? https://www.reddit.com/r/darksouls3/comments/727zfd/is_this_game_too_linear/. [Post Reddit].

SarcasticLizard (2016). Question about world design for those who played Bloodborne and Dark Souls 2, and how that will relate to Dark Souls 3. https://www.reddit.com/r/darksouls3/comments/430byz/question_about_world_design_for_those_who_played/. [Post Reddit].

SCP (2019). Is Dark Souls 2 linear ? https://www.reddit.com/r/DarkSouls2/comments/amdri8/is_dark_souls_2_linear/. [Post Reddit].

SJIS0122 (2023). Why was dark souls 3 so linear compared to the rest of the series ? https://www.reddit.com/r/fromsoftware/comments/16vl1ah/why_was_dark_souls_3_so_linear_compared_to_the/. [Post Reddit].

Triclot, M. (2012). Dedans, dehors et au milieu : les espaces du jeu vidéo. In Ter Minassian, H., Rufat, S., & Coavoux, S. (éd.), *Espaces et temps des jeux vidéo*, L>P. Questions théoriques, Paris.

Yu, D. & Hull, A. (2017). The Spelunky HD postmortem. https://www.youtube.com/watch?v=RiDy6CgBKqs. [Vidéo].

	Dark Souls	**Dark Souls II**	**Bloodborne**	**Dark Souls III**	**Elden Ring**
connecté ?	True	True	True	True	False
# de sommets	39	58	35	58	313
# d'arêtes	49	62	43	69	380
diamètre	12	22	10	24	45
excentricité moyenne	9.205	15.586	7.914	18.552	35.85
densité	0.066	0.038	0.072	0.042	0.008
fermeture triadique	0.157	0.122	0.132	0.256	0.178
distance moyenne	4.907	7.815	4.104	9.650	14.822
taille max. de clique	3	3	3	3	4
Cliques maximales	[3, 5, 9] [3, 5, 10] [3, 38, 9] [6, 9, 5] [24, 25, 27] [25, 26, 27]	[11, 12, 13] [12, 15, 13] [27, 28, 30] [40, 41, 39]	[2, 3, 4] [5, 6, 26] [5, 6, 13] [14, 13, 23]	[7, 8, 9] [9, 10, 47] [20, 21, 23] [21, 22, 23] [26, 27, 28] [31, 33, 34] [32, 33, 34] [37, 38, 44] [37, 38, 39] [40, 41, 42]	[33, 35, 40, 49] [33, 35, 40, 47] [77, 78, 80, 79] [131, 134, 133, 158]
Top degré des sommets	Firelink Shrine, 9 Darkroot Basin, 5 Anor Londo, 5 Undead Burg, 4 Undead Parish, 4	The Far Fire, 6 Servants' Quarters, 5 Cardinal Tower, 4 Ruined Fork Road, 4 Heide's Ruin, 3	Cathedral Ward, 8 Grand Cathedral, 5 Hemwick Charnel Lane, 4 1st Floor Sick Room, 3 Central Yharnam, 3	Church of Yorshka, 5 Dancer of the Boreal Valley, 4 Crucifixion Woods, 4 Farron Keep Perimeter, 4 Irithyll Dungeon, 4	Agheel Lake North, 8 Folly on the Lake, 7 Church of Elleh, 6 South of the Lookout Tower, 6 Academy Gate Town, 6
Top centralité de proximité	Firelink Shrine, 0.339 Undead Parish, 0.304 Blighttown Swamp, 0.299 Catacombs 1, 0.27 Undead Burg, 0.266	The Far Fire, 0.214 Old Akelarre, 0.204 Ruined Fork Road, 0.193 Undead Refuge, 0.191 The Crestfallen's Retreat, 0.188	Cathedral Ward, 0.386 Grand Cathedral, 0.378 Hemwick Charnel Lane, 0.34 Lecture Building, 0.318 Abandoned Old Workshop, 0.315	Crucifixion Woods, 0.15 Farron Keep Perimeter, 0.148 Halfway Fortress, 0.147 Road of Sacrifices, 0.143 Abyss Watchers, 0.143	East Capital Rampart, 0.092 Avenue Balcony, 0.089 Agheel Lake North, 0.089 Capital Rampart, 0.089 Gatefront, 0.087
Top centralité intermédiaire	Firelink Shrine, 0.731 Blighttown Swamp, 0.445 Undead Parish, 0.425 Sen's Fortress, 0.371 Anor Londo, 0.368	The Far Fire, 0.741 Ruined Fork Road, 0.541 Old Akelarre, 0.474 Undead Refuge, 0.338 Bridge Approach, 0.323	Cathedral Ward, 0.549 Grand Cathedral, 0.416 Lecture Building, 0.313 Lecture Building 2nd Floor, 0.236 Hemwick Charnel Lane, 0.214	Crucifixion Woods, 0.591 Catacombs of Carthus, 0.516 Farron Keep Perimeter, 0.509 Cliff Underside, 0.508 Halfway Fortress, 0.496	East Capital Rampart, 0.42 Grand Lift of Rold, 0.306 Agheel Lake North, 0.301 Forbidden Lands, 0.295 Stormhill Shack, 0.232

TABLEAU 1 – Tableau général des résultats pour les cinq jeux de FromSoftware.

Cahiers du CLSL, n° 69, 2025, pp. 239-254

Des mathématiques et des jeux. Le caractère ludique des graphes et des réseaux

Yannick Rochat
Université de Lausanne
yannick.rochat@unil.ch

Résumé

Dans ce chapitre, nous nous intéressons aux liens entre les mathématiques et divers types de jeux, en particulier ceux qui comportent une dimension relationnelle, ce qui invite à les étudier à l'aide de graphes et de réseaux. De nombreux jeux sont construits à partir de ces objets mathématiques, révélant ainsi leur caractère spontanément ludique. Ce chapitre propose un panorama de situations où se rencontrent jeux, graphes et réseaux.

Préambule

Les graphes et les réseaux[1] occupent une place centrale dans les recherches de notre collègue François Bavaud. François a pu constater l'importance croissante du jeu dans mes recherches lors d'un premier passage en section des sciences du langage et de l'information, entre 2016 et 2019. Je le remercie pour son généreux soutien lors de cette période charnière. Cet ouvrage collectif m'offre l'opportunité d'associer les graphes et les réseaux à ce qui est désormais mon principal objet d'étude. J'espère par ce texte te rassurer, cher François : je n'ai jamais tourné le dos à ces mathématiques que tu affectionnes tant.

1 Dans ce texte, nous considérons les réseaux comme des graphes dont les sommets et les arêtes possèdent une signification qui dépasse l'abstraction mathématique, étant le plus souvent ancrée dans le réel (liens entre personnes, lieux géographiques, proximité de sens entre entités, etc.).

1 Introduction

Quiconque a étudié la théorie des graphes et l'analyse de réseaux est en mesure de constater combien cette matière peut être approchée sous l'angle du jeu. Les graphes et les réseaux se dessinent facilement, se parcourent comme un plateau de jeu et peuvent même être transformés en puzzles à démêler. Ils offrent des opportunités amusantes pour explorer les relations au sein d'un groupe social, telle l'application en ligne *The Oracle of Bacon*[2] qui permet de trouver le chemin le plus court entre deux acteur·rices en passant par des tournages en commun avec d'autres acteur·rices. Les graphes et les réseaux jouent également un rôle central dans le fonctionnement du Web, ici à travers le concept des hyperliens. Le principe du jeu en ligne *The Wiki Game*[3] se base sur ceux-ci : en partant d'une page aléatoire de Wikipédia, l'objectif est de rejoindre une autre entrée aléatoire en n'utilisant que des hyperliens de l'encyclopédie. Le joueur ou la joueuse doit alors être capable de démêler la structure de l'encyclopédie en ligne pour obtenir les meilleurs scores.

Ce texte décrit des similarités et emprunts entre une catégorie d'objets mathématiques et un corpus d'œuvres ludiques dont la principale caractéristique est d'avoir des composantes interconnectées. Il débute par une discussion sur les jeux mathématiques, suivie d'une proposition de panorama des situations dans lesquelles jeux, graphes et réseaux se rencontrent. Cette esquisse de taxonomie inclut des exemples de systèmes de jeux construits sur des propriétés mathématiques, ainsi que des études de systèmes de jeux basées sur des graphes et des réseaux. Le texte se conclut par une discussion mêlant jeu et œuvres littéraires.

2 Mathématiques ludiques et jeux mathématiques

Qu'ils prennent la forme de jeux de société, de rôle, de cartes, de dés, ou de jeux vidéo, nombreux sont les jeux dont les systèmes, soit l'ensemble des mécaniques et des relations établies entre elles (Zubek, 2020, p. 5), reposent sur des propriétés mathématiques. Dans l'entrée *Mathématiques* du récent Dictionnaire des sciences du jeu, Lisa Rougetet retrace

2 https://oracleofbacon.org (consulté le 22 févr. 2025).
3 https://www.thewikigame.com/ (consulté le 22 févr. 2025).

des liens entre jeu et mathématiques remontant à l'Antiquité (Rougetet, 2024). Elle présente des exemples classiques de problèmes ayant mené à de nouveaux champs des mathématiques (Rougetet, 2024, p. 276), puis établit des parallèles entre l'activité mathématique et l'activité ludique. Résoudre un problème mathématique ou élaborer une stratégie au sein d'un jeu sont deux activités imposant de respecter un ensemble de règles, tout en manipulant divers objets : concepts mathématiques d'une part, pièces, dés, cartes voire plateau de jeu d'autre part.

Dans la suite de ce texte, nous ne traitons pas les « jeux mathématiques » tels qu'on les trouve par exemple dans l'œuvre de John Conway (Berlekamp et al., 2001; Conway, 2001), car ils s'éloignent de ce qui motive habituellement à jouer à des jeux. Le chercheur et *game designer* Robert Zubek écrit à propos de cette distinction :

> When game designers talk about the craft, there is an unspoken assumption that the games we make are for players to play. This sounds like an obvious truism, but not everybody treats games this way. Mathematicians, for example, may consider games as problems to be solved optimally, ignoring the player. (Zubek, 2020, p. 26)

Les jeux mathématiques sont d'abord des problèmes à résoudre, s'inscrivant dans la lignée des mathématiques récréatives[4]. Leur caractère ludique les destine le plus souvent à une audience particulière et restreinte, composée de mathématicien·ne·s professionnel·le·s et amateur·e·s. À ce sujet, la section *What is a game ?* de Berlekamp et al. (2001, pp. 14-16) propose une définition du *jeu* fondée sur huit critères stricts. Par exemple, un jeu doit comporter uniquement deux joueur·se·s (règle 1), les deux joueur·se·s doivent avoir accès à toute l'information (règle 5), et il ne doit pas y avoir de hasard (règle 6). Les auteurs examinent à la suite de cette définition comment certains jeux populaires, tels que le morpion, les échecs, le backgammon, la bataille navale [*Salvo* (1931)] ou le *Monopoly* (1935) [*The Landlord's Game* (1906)], ne respectent pas l'un ou l'autre de ces critères. Bien que d'autres définitions de jeu existent dans le domaine des mathématiques et que certains de ces

4 En référence au *Journal of Recreational Mathematics*, une revue scientifique qui a existé de 1968 à 2014. Ce n'est pas la seule revue à s'être consacrée à ce sujet.

« jeux », comme le *Hackenbush* (Berlekamp et al., 2001, p. 2), soient liés aux graphes et aux réseaux, nous excluons les jeux mathématiques du reste de ce texte.

3 Panorama

On fait généralement remonter la naissance de la théorie des graphes à l'article « *Solutio problematis ad geometriam situs pertinentis* » rédigé par Leonhard Euler il y a bientôt trois siècles (Euler, 1736). Dans ce texte, Euler offre la solution à un problème épineux de l'époque : déterminer s'il est possible de trouver un chemin passant une et une seule fois par les sept ponts reliant les quatre parties de la ville de Königsberg (aujourd'hui Kaliningrad). La réponse à cette question est négative. La résolution de ce problème, souvent présenté comme une énigme divertissante pour la population de cette ville, est considérée dans la majorité des manuels comme étant à l'origine de la théorie des graphes.

Or, parcourir un espace découpé en cases, dont certaines sont liées entre elles, invite à le modéliser sous forme de réseau. C'est par exemple le cas de la marelle, sans cesse recomposée au gré des lancers de cailloux, que les enfants parcourent à cloche-pied. Contrairement à d'autres objets mathématiques, l'idée de réseau est largement comprise par le grand public. Elle permet de percevoir rapidement les rôles et les propriétés des différents nœuds, ou acteurs, dans des jeux modélisant des relations, qu'elles soient sociales ou géopolitiques.

Le panorama qui suit vise à démontrer que les graphes et les réseaux possèdent un fort potentiel ludique en les étudiant selon trois fonctions distinctes : comme concepts de jeux, comme mécaniques de jeu, et comme outils d'analyse.

3.1 Des théorèmes aux jeux

Certaines propriétés mathématiques sont facilement transposables en jeux. L'un des exemples les plus célèbres est le *Dobble* (2009), un jeu composé d'un ensemble de cartes sur lesquelles apparaît un nombre fixe de symboles. La particularité de cet ensemble de cartes est que chaque paire de cartes partage toujours un et un seul symbole. Dans la

version standard, huit symboles sont affichés sur chaque carte, tirés d'un ensemble de 57 symboles, pour un total de 57 cartes[5]. Il s'agit d'une propriété liée à la géométrie projective (Bourrigan, 2011) et aux travaux de Thomas Kirkman (Lara, 2023) qui a énoncé et étudié le *problème des écolières* [*Schoolgirl problem*] :

> Fifteen young ladies in a school walk out three abreast for seven days in succession: it is required to arrange them daily, so that no two shall walk twice abreast. (Kirkman, 1850, p. 260)

On se trouve ici face à un cas où une énigme mathématique a conduit à des avancées dans ce domaine, donnant naissance à une situation dont on a tiré un jeu. Plus précisément, du matériel de jeu a été construit sur la base d'une propriété mathématique. Dans ce cas précis, les game designers laissent aux joueurs et joueuses la liberté de choisir leur variante parmi celles présentées dans le livret contenant les règles du jeu, ainsi qu'une notice historique. Ici, les mathématiques ont mené au jeu.

Nous observons maintenant quelques cas basés sur le thème qui nous intéresse ici, celui des graphes et des réseaux. Pour commencer, le jeu vidéo *Planarity*, publié en 2005 pour être utilisé dans un navigateur Web[6], illustre cette approche en invitant les joueurs et joueuses à déplacer les nœuds d'un graphe pour éviter que les arêtes ne s'entrecroisent (voir Fig. 1). Il est relativement simple de générer des puzzles pour ce jeu, consistant en des graphes planaires[7] disposés de manière peu optimale. En l'occurrence, la condition pour qu'un graphe soit planaire est qu'il ne contienne pas de sous-graphe complet d'ordre cinq (K_5), ni de sous-graphe biparti complet construit sur deux ensembles de trois sommets ($K_{3,3}$). Cet exemple met particulièrement en évidence l'intérêt de passer par un support numérique pour concevoir ce jeu.

5 Pour des raisons logistiques, le jeu ne comporte que 55 cartes.

6 Le site original ne permet plus de jouer à ce jeu vidéo. On peut en trouver une version équivalente ici : `https://www.chiark.greenend.org.uk/~sgtatham/puzzles/js/untangle.html` (consultée le 23 févr. 2025).

7 Un graphe est planaire s'il est possible de le représenter visuellement sans que deux arêtes ne s'intersectent.

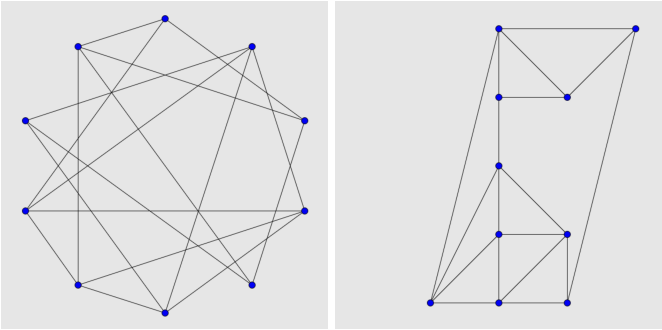


FIGURE 1 – Un graphe planaire composé de 10 sommets et de 18 arêtes. À gauche, tel que le jeu *Planarity* nous le présente. À droite, le même graphe après résolution : cette représentation ne comporte aucune intersection d'arêtes.

Plus ancien, le *Icosian game* est un jeu créé par le mathématicien William Rowan Hamilton et commercialisé en 1856 (Ashlock, 2016, p. 248). Il se joue sur un plateau représentant un dodécaèdre aplati correspondant à un graphe à 20 sommets où chaque sommet (troué dans le plateau) est connecté à trois autres sommets. L'objectif est de placer vingt pions numérotés dans tous les trous du plateau avant de revenir au point de départ, formant ainsi un cycle hamiltonien puisque chaque sommet est visité une et une seule fois. Le jeu n'a pas rencontré le succès, probablement parce qu'il était intimidant de prime abord et trop facile une fois les règles comprises (Ashlock, 2016, p. 249). Cependant, le plateau du *Icosian game* a pu servir d'inspiration pour d'autres jeux et énigmes liés aux graphes (Ashlock, 2016, p. 252). Notons qu'il s'agit d'un jeu de société conçu par un mathématicien célèbre.

En 1852, Francis Guthrie conjecture qu'il est possible de colorier une carte géographique arbitraire à l'aide de quatre couleurs au maximum de sorte que deux régions adjacentes ne partagent jamais la même couleur (Fritsch & Fritsch, 1998, p. 1). Ce résultat, prouvé de manière computationnelle dans la deuxième moitié du vingtième siècle seulement, est connu aujourd'hui comme le théorème des quatre couleurs. C'est le point de départ de divers jeux, dont le jeu de société *Tower Up* (2024), qui invite à construire des gratte-ciels sur une carte représentée par un

FIGURE 2 – Le plateau de jeu de *Tower Up*.

réseaux, de manière à ce que deux gratte-ciels adjacents ne soient jamais de la même couleur (voir Fig. 2). Plus que *Dobble* et *Planarity*, *Tower Up* a nécessité un travail important en matière de game design avant d'atteindre sa forme actuelle, trouvant un équilibre entre une propriété mathématique à exploiter et des mécaniques d'interaction entre joueurs et joueuses rendant chaque décision déterminante.

Des connaissances mathématiques peuvent être adaptées sous la forme de jeux. En fonction des ambitions des game designers, ces propositions nécessiteront ensuite plus ou moins d'efforts pour être transposées en expériences de jeux partageables avec un public étendu.

3.2 Des graphes aux jeux

Un jeu peut aussi recourir à une structure de graphe sans reposer sur un théorème précis. Dans ce cas, plutôt que de partir d'un résultat pour en déduire un système de jeu et construire ensuite un ensemble de règles, on décide de créer un jeu possédant une dimension relationnelle. Il s'agira par exemple de déplacements de pions ou de connexions entre des lieux.

Certains systèmes de jeux reposent sur des structures en réseau non explicitées. C'est le cas des échecs, qui se jouent sur un quadrillage de 64 cases. Bien que l'on puisse oublier le maillage sous-jacent, celui-ci, associé aux règles de déplacement spécifiques de chaque type de pièce, guide les joueurs et joueuses. Ce maillage de l'échiquier a inspiré de nombreux détournements mathématiques, dont le célèbre problème du cavalier, qui consiste à faire parcourir toutes les cases de l'échiquier par un cheval, une et une seule fois (Rittaud, 2015).

Les jeux basés sur des maillages réguliers ne sont pas rares, comme en témoigne le long héritage des *wargames*, un genre vieux de plusieurs siècles modélisant des conflits militaires à l'aide de plateaux composés de cases le plus souvent carrées ou hexagonales, la plupart du temps sans vide entre elles (von Hilgers, 2012). Dans le cas de jeux de société plus accessibles, la présence d'un territoire arpentable sur le plateau de jeu, souvent associé à un contexte géopolitique, augmente nettement la probabilité que la «structure » sous-jacente d'un système de jeu donné soit un graphe. Par exemple, le célèbre *Risk* [*La Conquête du Monde*, (1957)] est une simulation militaire où la résolution des conflits repose sur des tirages de dés à six faces. Le déroulement de la partie dépend fortement de probabilités simples à calculer[8]. Relativement ancien, il ne comporte pas visuellement de réseau (à l'exception des liaisons intercontinentales), là où les jeux plus récents n'hésitent pas à expliciter les relations entre deux zones limitrophes par une arête plutôt que par une frontière commune. On peut d'ailleurs se demander si cette présence accrue des réseaux dans les jeux contemporains résulte d'une quête de lisibilité éditoriale, ou dérive plutôt d'une familiarité culturelle (voire d'une reconnaissance esthétique) du motif du réseau auprès du grand public (Jagoda, 2016).

Autre exemple célèbre, le jeu de société *Catan* (1995) invite joueurs et joueuses à étendre leur domaine sur un plateau aux cases hexagonales pour récupérer des ressources nécessaires à leur expansion (voir Fig. 3). Le bon déroulement du jeu repose sur le placement stratégique de villes

8 C'est le cas lors d'un affrontement, ce qui ne garantit pas que le résultat soit celui attendu. Notons que les mathématiciens Daniel Ashlock et Colin Lee ont calculé le positionnement stratégique de chaque région de la carte à l'aide d'une méthode d'estimation de diffusion de gaz et de mesures d'entropie (Ashlock & Lee, 2015).

FIGURE 3 – Le plateau de jeu d'une version étendue de *Catan*. Des pions représentant les villes et les villages sont placés aux sommets des hexagones. Les villes et les villages sont reliés par des bâtons symbolisant des routes. Photographie de Yonghokim (CC BY-SA 4.0).

et villages sur des nœuds à fort rendement, ainsi que sur le positionnement de routes le long des arêtes les plus prometteuses. Le tirage des dés détermine les ressources obtenues : un bon positionnement augmente les chances de succès, mais le hasard et les stratégies des adversaires, qui cherchent à s'étendre avant nous, offrent un équilibre entre stratégie, bluff et chance, contribuant ainsi à l'immense succès du titre auprès d'un large public[9].

Plus simple et plus élégant, *TransAmerica* (2001) se déroule sur un plateau relativement austère, doté d'un maillage régulier et triangulaire. La boucle de jeu consiste à placer, à chaque tour, une ou deux voies sur ce maillage. L'objectif est de concevoir en premier un arbre[10] reliant cinq villes de la carte, tirées aléatoirement en début de partie et gardées cachées par chaque participant·e. Lorsque les réseaux de deux personnes se rejoignent, ils n'en forment plus qu'un, facilitant soudainement les liaisons avec des points éloignés. Le jeu propose ainsi de construire collectivement un arbre couvrant de poids (relativement) minimal dans l'espoir qu'il nous avantagera par rapport aux autres. Cette forme de

9 Pour une analyse avancée de *Catan*, voir Guhe & Lascarides (2014).

10 Aucune règle n'oblige à construire un arbre, soit un graphe sans cycle, mais poser une arête formant un cycle n'améliorera pas la situation du joueur ou de la joueuse.

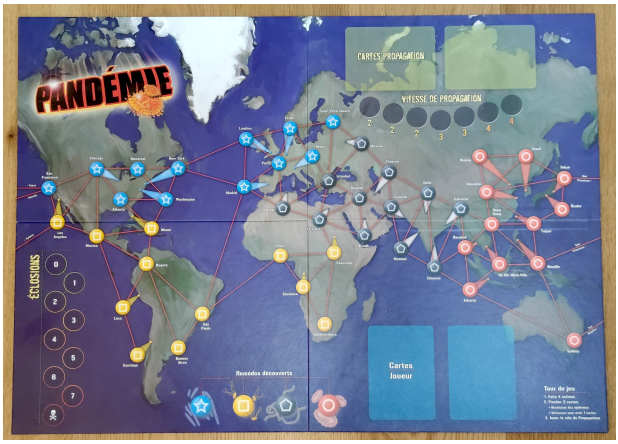


FIGURE 4 – Le plateau du jeu *Pandémie* [*Pandemic*], 2008.

coopération, qui rappelle la théorie des jeux, implique d'anticiper et d'exploiter les décisions des autres. Une idée simple, basée sur la théorie des graphes et assortie d'une dimension d'incertitude absente des jeux mathématiques, peut être un facteur essentiel pour concevoir un jeu de société réussi.

Le jeu de société collaboratif *Pandemic* en est l'illustration (voir Fig. 4). Celui-ci attribue des rôles tous différents aux joueur·ses et leur demande ensuite tour à tour de se répartir les tâches à mener dans le monde pour contenir, voire éradiquer une série de quatre (!) pandémies. Parmi ces tâches, il faut développer des vaccins en laboratoire et lutter contre les maladies sur le terrain. Le jeu demande de se déplacer, sur terre ou par les airs, et une réévaluation constante de la situation doit être effectuée pour s'assurer que les points d'action à disposition ne soient pas gaspillés. Dans la pratique, *Pandemic* (2008) invite chaque personne confrontée à ce réseau représentant le monde à assimiler la notion de plus court chemin (en tout cas dans un contexte ludique). En revanche, le jeu se complexifie significativement sitôt que les actions à mener lors d'un même tour se multiplient. Les partenaires élaborent une stratégie collective pour appréhender efficacement le réseau, une stratégie qui demandera probablement d'être renégociée de nombreuses

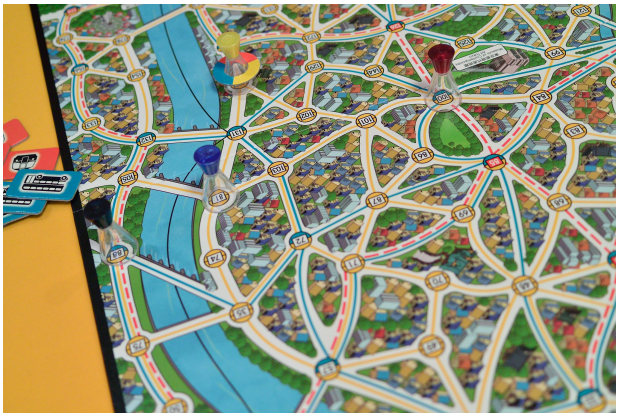


FIGURE 5 – Une portion du plateau de jeu de *Scotland Yard : Tokyo* (2014). Photographie de yoppy (CC BY 2.0).

fois en cours de jeu. On voit donc que le réseau, qui figure les villes infectées ou prêtes à l'être, est à la fois un support du jeu et un facteur de victoire, cette dernière n'étant offerte qu'à une équipe qui aura acquis une compréhension suffisante du réseau.

La transposition de certaines infrastructures vers le jeu est un terrain privilégié pour l'apparition de réseaux ou pour solliciter les mécaniques qu'ils peuvent engendrer. Dans *Scotland Yard* (1983), un·e joueur·se tente de s'enfuir à travers la ville de Londres, tandis que les adversaires sont à sa poursuite mais ne possèdent que des informations partielles sur sa position (voir Fig. 5). Pour s'enfuir à travers la ville (modélisée par un réseau), il est permis d'utiliser le taxi, le bus, le métro, voire le bateau, pour des coûts divers. Le jeu invite à parcourir un réseau composé des trajectoires potentielles liées à ces moyens de transport. Dans ce cas, où le jeu est asymétrique, le réseau est exploité par une personne pour ses déplacements et évalué par les autres, qui doivent alors calculer toutes les possibilités de déplacement de la personne en fuite et deviner laquelle est la bonne, ce qui peut mener à des triangulations victorieuses ou, au contraire, à ce que la personne passe à travers les mailles du filet et s'échappe. Le réseau est le terrain de jeu d'une partie élaborée de cache-cache.

FIGURE 6 – Le plateau de l'extension « Japon » pour le jeu *Power Grid* [*Funkenschlag*], 2010.

Autre jeu présentant une infrastructure, cette fois industrielle, *Funkenschlag* (2001) propose de construire des centrales électriques et de gérer l'approvisionnement d'un réseau de villes (voir Fig. 6). Ce type de gameplay se retrouve également dans le jeu de société à succès *Ticket to Ride* (2004), qui propose quant à lui de (re)construire un réseau ferroviaire en suivant des traces existantes et mobilise l'attrait pour l'exotisme en déclinant ses terrains de jeu dans divers contextes géographiques (voir Fig. 7). Dans ces différents cas, on parcourt un réseau justifié par un contexte thématique, où les arêtes correspondent à des éléments physiques, sans prendre de pincettes pour expliciter graphiquement le réseau, qui a en réalité été conçu et renégocié lors de longues phases de tests (Zagal, 2023).

En conclusion, l'intégration des graphes et des réseaux dans les jeux de société constitue une solution efficace pour gérer des ressources et des dynamiques territoriales au sein de nombreux systèmes de jeu. Des concepts issus de la théorie des graphes et de l'analyse des réseaux, tels que la centralité, la recherche des plus courts chemins et l'optimisation d'arbres couvrants de poids minimal, peuvent être exploités pour concevoir des mécaniques de jeu engageantes. Dans le cadre des jeux de société basés sur des graphes, la quête d'un game design réussi réside dans l'harmonisation de ces mécaniques avec la cohérence thématique.

FIGURE 7 – Le plateau de jeu de *Ticket to Ride*, version Europe.

3.3 Étudier des systèmes

Si les réseaux peuvent servir de substrat principal allant jusqu'à constituer un plateau de jeu, le recours à ceux-ci est souvent pertinent pour modéliser les états d'un jeu et les étudier, quand bien même le concept de graphe n'est pas au centre des mécaniques ludiques (Ashlock, 2016, p. 255). Ces méthodes peuvent être déployées dans les jeux où il s'agit de parcourir un plateau avec une probabilité associée de passer d'une case à une autre. C'est le cas par exemple dans le jeu des serpents et des échelles, le jeu de l'oie ou encore les échecs. Il est alors possible de modéliser tous les déplacements à l'aide d'un graphe, en associant une probabilité à chaque arc (arête dirigée), et d'en déduire quelle stratégie privilégier dans le cas où le joueur ou la joueuse prend activement des décisions.

Le mathématicien Daniel Ashlock synthétise la rencontre entre jeux, graphes et réseaux, en plaçant au même niveau game designers et stratèges amateurs dans un article intitulé *Graph Theory in Game and Puzzle Design* :

> An understanding of basic graph theory permits game designers, or anyone interested in the algorithmic solution of puzzles, to abstract the essential features of boards, solution spaces, or even rule sets. The abstraction gives a different,

> more concise, view of some aspect of a game. It may not be a better [instrument] for addressing a particular design issue, but is an additional tool for the designer. (Ashlock, 2016, p. 255)

L'étude du *Monopoly* est un sujet relativement fréquent dans ce domaine (Bernard, 2017; Fry & Evans, 2016; Stewart, 2004), révélant l'intérêt de se baser sur la culture populaire pour inviter à découvrir les mathématiques (et vice-versa). Ces études font appel aux chaînes de Markov, très appréciées pour modéliser les propriétés de divers systèmes de jeux. Elles permettent d'identifier les meilleures stratégies en tenant compte des contextes changeants de l'environnement de jeu : achat de terrains, construction de maisons et d'hôtels, paiement de loyers, etc.

3.4 Graphes, littérature et jeu

Ce panorama ne saurait omettre certains mouvements artistiques qui ont explicité les liens entre mathématiques, mécaniques ludiques et textualités[11]. À cet égard, l'OuLiPo fait figure d'incontournable. Né dans les années 60 en France et mêlant des adeptes de mathématiques comme de littérature, le collectif est à l'origine de curiosités et de chefs-d'œuvre. *La vie mode d'emploi*, rédigé par Georges Perec et publié en 1978, invite à découvrir l'intimité des habitant·e·s d'un immeuble. Très présent dans le récit, le jeu ne fait pas qu'apparaître sous la forme de puzzles. Perec choisit en effet de calquer l'ordre de présentation des personnages sur une solution du problème du cavalier déployée sur le plan de l'immeuble, dont les appartements forment l'échiquier. Le récit dépend dès lors d'un contexte ludique autant que mathématique.

Pour qui n'aurait pas la patience de découvrir l'ample roman de Perec, l'énigme sous forme de nouvelle *Qui a tué le duc de Densmore ?* du mathématicien français Claude Berge offre un exemple tout aussi intéressant de croisement entre littérature et mathématiques. Publié en 1990, ce récit d'enquête nécessite de tirer parti des informations fournies afin de construire le graphe qui permettra d'établir l'identité du ou de la coupable (Berkman, 2022, pp. 218-219). Preuve supplémentaire du fait que littérature et mathématiques gagneraient à se fréquenter davantage.

11 À ce propos, voir le chapitre d'Isaac Pante présent dans cet ouvrage.

4 Conclusion

Dans cette contribution, nous avons passé en revue diverses manières de *jouer* avec des graphes et des réseaux. Dans ce contexte, les usages possibles de ces derniers sont riches. Ils peuvent se révéler à la fois intuitifs et d'une grande complexité, et occupent une place privilégiée dans le cours de *Mathématiques ludiques*, qui fait partie du plan d'étude d'informatique pour les sciences humaines à l'Université de Lausanne. Ce cours a notamment pour objectif de familiariser les étudiant·e·s avec l'étude du jeu ainsi qu'avec divers thèmes mathématiques lors de leur cursus en lettres.

> Like games and puzzles, graph theory is a natural path to reducing students' anxiety about mathematics and computer science. (Ashlock, 2016, p. 256)

Une tâche qui se voit facilitée par la manipulation des graphes et des réseaux en cours.

Références

Ashlock, D. (2016). Graph theory in game and puzzle design. *Game & Puzzle Design*, 2(1):62–70.

Ashlock, D. & Lee, C. (2015). Influence maps and new versions of risk. *Game & Puzzle Design*, 1(1):38–43.

Berkman, N. (2022). *OuLiPo and the mathematics of literature*. Number 141 in Modern French Identities. Peter Lang, Oxford, 1ère édition.

Berlekamp, E. R., Conway, J. H., & Guy, R. K. (2001). *Winning ways for your mathematical plays: Volume 1*. A K Peters/CRC Press, New York, 2ème édition.

Bernard, B. (2017). Monopoly – An Analysis using Markov Chains. https://carlabernard.ch/beni/downloads/bernard_monopoly.pdf. [PowerPoint slides].

Bourrigan, M. (2011). Dobble et la géométrie finie. *Images des mathématiques*. https://images-archive.math.cnrs.fr/Dobble-et-la-geometrie-finie.html.

Conway, J. H. (2001). *On numbers and games*. A.K. Peters, Natick, MA, 2ème édition.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Opera Omnia*, 7:128–140.

Fritsch, R. & Fritsch, G. (1998). *The Four-Color Theorem: History, topological foundations, and idea of proof*. Springer, New York.

Fry, H. & Evans, T. O. (2016). *The indisputable existence of Santa Claus: The mathematics of Christmas*. Doubleday, London.

Guhe, M. & Lascarides, A. (2014). Game strategies for the settlers of Catan. In *2014 IEEE Conference on Computational Intelligence and Games*, pages 1–8, Dortmund, Germany. IEEE.

von Hilgers, P. (2012). *War games: A history of war on paper*. MIT Press, Cambridge, MA.

Jagoda, P. (2016). *Network aesthetics*. University of Chicago Press, Chicago, London.

Kirkman, T. P. (1850). Note on an unanswered prize question. *Cambridge and Dublin Mathematical Journal*, 5:255–262.

Lara, J. S. (2023). Spotting k-TriCaps in SPOT IT! Technical report, Bard College, Hudson, NY.

Rittaud, B. (2015). Graphes sur échiquier. *Les graphes : représenter les données et les stratégies*, HS 54:148–150.

Rougetet, L. (2024). Mathématiques. In Brougère, G. & Savignac, E. (éd.), *Dictionnaire des sciences du jeu*, Questions de société, pages 276–281. Érès, Toulouse.

Stewart, I. (2004). *Math hysteria: Fun and games with mathematics*. Oxford University Press, Oxford.

Zagal, J. P. (2023). Pandemic: When the abstract becomes concrete. In Randl, C. & Lasansky, D. M. (éd.), *Playing place. board games, popular culture, space*, pages 127–129. The MIT Press, Cambridge, MA.

Zubek, R. (2020). *Elements of game design*. The MIT Press, Cambridge, MA.

Les réseaux des travaux de François Bavaud dans tous leurs étaux ou comment aborder François Bavaud dans le style de « Cantatrix Sopranica »

Céline Rozenblat & Mikhail Rogov
Université de Lausanne
{celine.rozenblat,mikhail.rogov}@unil.ch

Résumé

Cet article rend hommage à François Bavaud en retraçant les grandes lignes de son parcours scientifique, marqué par une approche interdisciplinaire alliant géographie quantitative, analyse textuelle, physique statistique et théorie de l'information. Inspirés de la méthode ludique de Perec, les auteurs explorent ses travaux à travers une analyse textuelle de ses publications, de leurs sources et de leurs citations, mettant en lumière des réseaux conceptuels complexes et originaux. Trois articles clés (Bavaud, 1991, 1998, 2011) servent de pivots à cette exploration : le premier sur la mécanique statistique, le second sur les matrices spatiales pondérées, et le dernier sur les transformations de Schoenberg. Chacun illustre la manière dont François Bavaud tisse des liens entre différentes disciplines et construit des ponts théoriques et méthodologiques. Son œuvre est saluée pour sa liberté intellectuelle, sa créativité mathématique et sa contribution à des domaines aussi variés que la physique, la géographie, l'écologie ou le machine learning. L'article souligne enfin l'héritage scientifique qu'il lègue à des communautés multiples, à travers une pensée rigoureuse, indépendante et profondément singulière.

Préambule

Tout le monde connaît François Bavaud, et pourtant qui le connaît vraiment ? Défenseur de la géographie quantitative (mais non inféodé), François Bavaud est tout autant friand d'analyses textuelles à la recherche du sens retrouvé des textes, films ou musiques. À cheval sur deux facultés, la Faculté des Géosciences et de l'environnement et la

Faculté des Lettres, il trace depuis de nombreuses années son chemin scientifique en statistiques spatiales et textuelles de manière libre et indépendante. « Libre », c'est un qualificatif qui, nous l'espérons, lui plaira, et qui nous semble bien correspondre à sa personnalité. Toutefois, on peut se demander s'il est scientifiquement si libre que cela. N'est-il pas finalement « coincé », « pris en étau », entre des paradigmes forts qu'il se doit de respecter, pour à son tour les léguer à une postérité qui s'en inspirera ? Et si oui, quels sont ces paradigmes, qu'en a-t-il fait et qu'en font ses héritiers ?

C'est ce que nous nous sommes demandé à ce moment précis où nous célébrons cette étape importante de sa vie professionnelle, en prenant au mot le terme « retraite » pour re-traiter ses productions scientifiques. Pour cela, nous avons considéré François Bavaud comme un objet d'étude que nous avons analysé à travers son corpus de publications, d'une manière dont on nous pardonnera la très modeste méthodologie au regard de la sophistication statistique de ses propres recherches. Néanmoins, nous nous sommes pris au jeu de l'exercice en appliquant une méthodologie que nous avons voulue rigoureuse et systématique.

Inspirée de « Cantatrix sopranica » (Georges Perec, 1991)[1], sans prétendre nullement au génie chaotique de Perec, l'approche insistera sur les associations de termes, autant pour souligner la richesse qui apparaît dans les contributions scientifiques de François Bavaud, que pour interroger leur diversité, qui peut parfois paraître, elle aussi, chaotique au point de maximiser l'entropie du propre système théorique de l'auteur.

1 Approche et méthodologie biographique

Afin d'explorer le paysage conceptuel des travaux de François Bavaud, nous avons commencé par extraire un corpus de ses publications à partir de Google Scholar, soit 65 articles référencés. Nous avons considéré les résumés de ces articles afin d'en extraire le vocabulaire. Nous avons construit les cooccurrences des mots et la carte sémantique de

1 Dans cette perspective, une fausse référence se cache dans les nombreuses vraies références… et ce n'est pas Caussinus qui est une vraie référence. Egalement pour rendre hommage au regretté Henri Chamussy, vous trouverez (ou non) trois contrepèteries au fil du texte.

ces réseaux. Chaque nœud du réseau correspond à un terme significatif, tandis que les arêtes illustrent la fréquence à laquelle les termes apparaissent dans les mêmes résumés. Emergent des grappes qui indiquent les associations des sujets dominants et l'organisation des modèles intellectuels. Cette méthode permet d'identifier les multiples contributions à différents domaines, leur diversité et leur cohérence.

Pour approfondir l'analyse de l'influence scientifique et du contexte intellectuel de François Bavaud, nous avons ensuite sélectionné trois de ses publications clés qui sont les plus citées jusqu'à présent (Bavaud, 1991, 1998, 2011) et nous avons construit deux corpus distincts pour chacune d'entre elles : l'un comprenant toutes les références incluses dans cette publication, et l'autre avec tous les articles qui citent la publication cible. Nous avons ainsi obtenu six corpus textuels, chaque paire de corpus associée à l'un des trois articles de François Bavaud représentant les dialogues scientifiques entrants et sortants. Dans les six graphes chronologiques en amont et en aval, les nœuds représentent les publications et les arêtes révèlent les relations de citation. Nous avons analysé et visualisé les champs conceptuels de ces corpus en cartographiant également à chaque fois la cooccurrence de leurs termes clés (six graphes également).

Cette approche va permettre une exploration comparative de la façon dont le travail de François Bavaud s'appuie sur des idées fondamentales et, d'un autre côté, comment il est reçu et interprété par des travaux ultérieurs. Cette double perspective permet de retracer visuellement la lignée savante de ses recherches, en cartographiant les auteurs et les idées sur lesquels il s'est appuyé, ainsi que la trajectoire de son influence à travers les auteurs qui se sont à leur tour servis de son travail. En plaçant les articles de François Bavaud au centre de chaque réseau, nous avons produit une vision temporelle de la manière dont ses recherches se situent dans le discours académique plus large, révélant les fondements intellectuels sur lesquels il s'est appuyé et les diverses manières dont ses contributions ont nourri des travaux ultérieurs.

Les réseaux sémantiques qui en résultent donnent un aperçu des passerelles entre le cadre conceptuel de Bavaud, les influences intellectuelles en amont et les communautés universitaires qui, en aval, se

sont appuyées sur son travail. La dynamique des citations de ces trois articles est toutefois à prendre avec prudence car, comme nous le verrons, ses travaux sont variés et articulent simultanément des champs différents, sans forcément qu'il y ait de succession préférentielle clairement identifiable dans le temps, sans qu'il en ait changé les maths. Dans un souci de cohérence dans cette dernière partie sur le corpus des articles qui citent François Bavaud, nous n'évoquerons pas les articles d'autocitation, bien qu'ils soient compris dans le corpus.

D'un point de vue technique, le travail sur tous ces corpus a commencé par une sélection des mots pour dégager les principaux termes qui peuvent donner du sens et éventuellement un regroupement des expressions similaires. Les cooccurrences des termes dans les mêmes articles ont permis de les positionner relativement les uns aux autres grâce au layout « Force Atlas ». Des clustering de graphes de « Louvain » (Blondel et al., 2008) ont attribué les termes dans des classes les plus cohérentes possibles (couleur) que nous avons tenté de nommer. Les citations et références des trois articles de Bavaud ont été collectées avec l'aide de l'outil Litmaps. Pour réaliser l'analyse textuelle et réaliser les graphes, nous nous sommes appuyés sur les logiciels libres Gargantext (Delanoë et al., 2023) et Gephi[2].

2 Les réseaux conceptuels de François Bavaud

Les concepts et méthodes clés développés dans les 65 publications de François Bavaud que nous avons analysées forment des paquets de champs conceptuels qu'il n'est pas aisé de qualifier clairement tant les concepts sont bien souvent combinés (Fig. 1).

Au centre du graphe, on trouve les formalismes fondamentaux dans les classes nommées ici **Classification properties** et **Matrix similarities**.

Matrix similarities regroupe des méthodes aussi diverses que les *chaînes de Markov* ou des *indices de similarité* ou de *concentrations locales*. Les *Markov chains* (*reversible Markov transition matrices*) permettent, par exemple dans le champ textuel, de quantifier des cooccurrences itérées pour construire des *similarity indices* entre mots ou

2 https://gephi.org/ (consulté le 13 mai 2025).

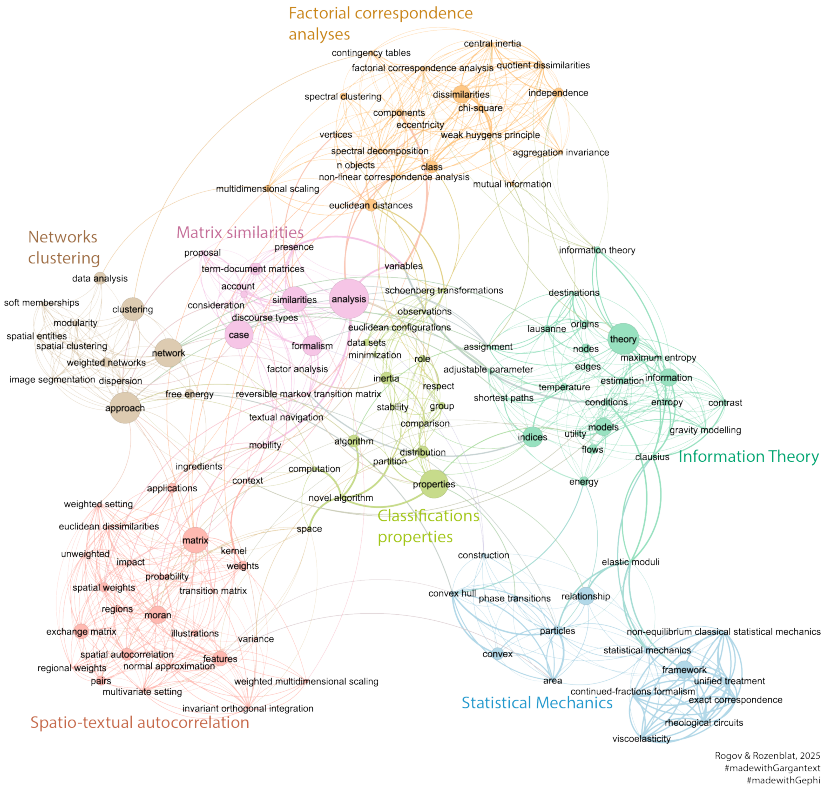


FIGURE 1 – Principaux concepts des articles de François Bavaud (1986-2024).

entre documents et de les classifier à différents ordres de cooccurrence (Bavaud & Xanthos, 2005). Les indices de similarité sont également appliqués aux matrices *terme-document* pour tenir compte des similarités sémantiques entre les termes, contribuant à réduire la variété du contenu dans des *correspondence analyses* (Egloff & Bavaud, 2018). Ces indices de similarité sont également utilisés pour comparer des *fuzzy classifications* (Bavaud, 2004) ou pour proposer un nouvel *indice de diversité*, basé sur l'*effective entropy* qui, selon l'auteur, réduit l'entropie de Shannon en tenant compte de la présence de similitudes entre les éléments (Bavaud, 2022).

Le groupe **Classification properties** est également assez central,

dans la mesure où il regroupe des termes généraux employés dans un grand nombre de ses articles : *observations*, *properties*, *algorithms*, *inertia* et *distributions* sont au cœur des réflexions sur le *partitionnement* où François Bavaud met en perspective la définition des distances euclidiennes sur les *graphes pondérés* et leurs *propriétés*. Cette réflexion est particulièrement développée dans trois publications entre 2010 et 2011 : partant des « Euclidean distances, soft and spectral clustering on weighted graphs » (Bavaud, 2010a), François Bavaud s'attache alors à approfondir les transformations de Schoenberg (Bavaud, 2010b, 2011). Grâce à ces transformations de Schoenberg (1938), il analyse les propriétés de nouvelles estimations de localisation qui transforment les distances euclidiennes initiales en nouvelles distances euclidiennes.

On reste dans des méthodes générales en glissant vers la classe de termes **Statistical mechanics** qui se réfère à des travaux précoces de la carrière de François Bavaud en Mécanique statistique portant sur les propriétés des formes convexes, les *elastic moduli* et la *viscoelasticity* pour la circulation des fluides (*rheology*) (Bavaud, 1987, 1989; Bavaud et al., 1986).

Ce qui permet d'intégrer la mécanique statistique aux problématiques actuelles se trouve autour des notions de *flow (origin-destination)*, d'*information theory*, d'*energy*, de *temperature*, de *thermodynamics* et d'*entropy*, qui sont des termes associés au groupe voisin que nous avons intitulé **Information theory**. Ce groupe de termes représente une partie théorique des travaux, où il unifie l'énergie, l'entropie et la théorie de l'information, les précise dans Zipf et Pareto, ce qui le conduit à tester l'effet de la variation de température et de « chauffer ou refroidir des textes » (Bavaud, 2009, p. 74 ; Bavaud & Xanthos, 2002, p. 4). Expliquant que finalement tout ceci repose sur la loi du moindre effort, le *gravity modelling* et les *shortest paths* sont étudiés dans le cadre de la *théorie de l'Information* qu'il semble préférer à toute autre.

L'origine des liens entre toutes ces théories, qu'il explique parfois à demi-mots dans ses articles, vient de la cristallographie dont sont issus d'autres collègues de renommée internationale dans la physique des réseaux (Havlin & Ben-Avraham, 1987). Ces racines de François Bavaud demeurent prégnantes dans ses travaux d'application qu'il a pu

faire tant en géographie qu'en analyses textuelles. C'est là un premier résultat majeur : François Bavaud se plaît à rapprocher conceptuellement les théories successives ou simultanées en associant allègrement différentes approches et à les faire dialoguer jusqu'à aujourd'hui.

C'est ainsi que toute cette base définit aussi son approche originale que l'on trouve dans les trois groupes de termes restants : **Factorial correspondence analyses**, **Network clustering** et **Spatio-temporal autocorrelation**. Ces trois groupes d'approches ont en effet en commun la théorie de l'information, les matrices de similarités et les propriétés des classifications, des partitions et plus fondamentalement la mécanique statistique. A ce stade, nous espérons avoir éclairé le·la lecteur·rice et ne pas l'avoir perdu·e.

3 Les sources de François Bavaud

C'est pourquoi nous proposons de remonter aux sources de cette œuvre. Ces sources, nous les avons puisées dans trois articles emblématiques de trois périodes de ses écrits : 1991, 1998, 2011.

3.1 Des sources initiales issues de « statistical mechanics »

Le premier article est :

> Bavaud, F. (1991). Equilibrium properties of the Vlasov functional: the generalized Poisson-Boltzmann-Emden equation. *Reviews of Modern Physics*, 63(1):129–149.

Dans cet article, François Bavaud affirme

> New results concerning the thermodynamic limit, phase transitions, metastability, and the shape of density profiles are provided. In particular, the question of ground states (in relationship to condensation and wetting phenomena) is illustrated by numerous explicit solutions (Bavaud, 1991, p. 129).

Il aborde des concepts aussi variés et originaux pour les néophytes que « superharmonicity », « degenerate ground state », « rotational invariance of density profiles » ou « truncation of higher-order correlations », soit un grand écart courageux de formulations de concepts basés sur une bibliographie comptant pas moins de 95 références (Fig. 2).

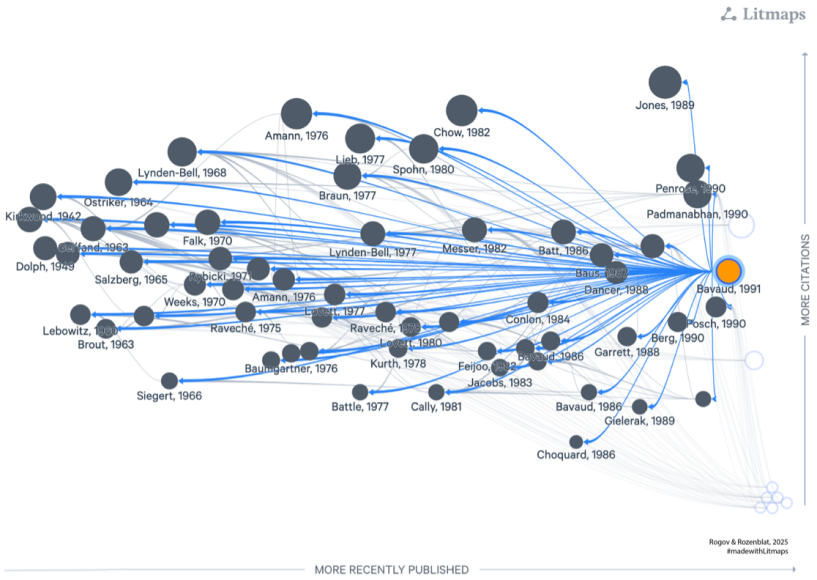


FIGURE 2 – Références bibliographiques de l'article Bavaud (1991).

Ces références couvrent un large champ de *Statistical mechanics* (Fig. 3). L'une traite de **Change of state** par *bifurcation* (Raveché & Stuart, 1976), par *freezing*, (Raveché & Kayser, 1978), avec la célèbre *first BBGKY equation* (pour les initiales de : Bogolioubov, Born, Green, Kirkwood et Yvon) ou par *fusion* (Kirkwood & Monroe, 1941). Ces systèmes concernent autant des *star systems* que des molécules (*fermions*, *bosons*) que l'on trouve dans les classes **Equations**, **Functions** et **Energy**. S'isolent les *Coulomb systems* (Conlon, 1984; Kennedy, 1984; Kiessling, 1990) (classe nommée **Statistical mechanics**) associés à *equilibrium thermodynamics*, *thermodynamic limits* ou *gravitational phase transitions* (Baumgartner, 1976).

3.2 Rencontre avec les statistiques spatiales

Le deuxième article que nous analysons est :

> Bavaud, F. (1998). Models for spatial weights: a systematic look. *Geographical Analysis*, 30(2):153–171.

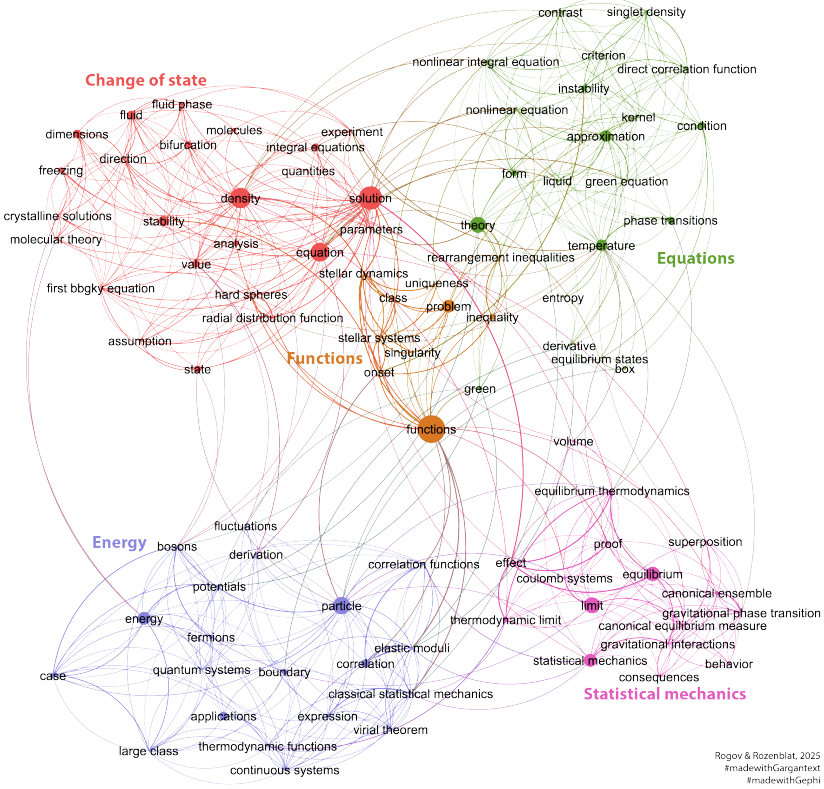


FIGURE 3 – Thématiques des références de l'article Bavaud (1991).

A la parution de cet article, François Bavaud vient de prendre son poste de professeur assistant de modélisation mathématique et de statistique à l'Université de Lausanne, car il est profondément persuadé que c'est un beau métier, professeur. Il revient d'un post-doc à l'Institute for Mathematical Behavioral Sciences – University of California – Irvine, où il a commencé l'écriture du manuscrit. Cette mobilité lui a permis de découvrir, s'approprier et développer les statistiques spatiales. Il s'intéresse particulièrement à la distribution stationnaire associée aux poids spatiaux (l'indice de proéminence d'import ou d'export), aux classes structurelles d'interaction et aux propriétés d'invariance par inversion

temporelle ou par agrégation des matrices avec des poids spatiaux. Il les applique aux matrices de transition des chaînes de Markov, ce que l'on retrouve dans les modèles spatiaux, mais surtout il dit viser à combler une lacune de manque de géométrie intégrale dans ces approches. Il utilise neuf exemples, impliquant la connectivité, les flux et les modèles de décroissance de la distance, la géométrie intégrale et les tessellations de Dirichlet-Voronoï, pour illustrer la démonstration de ces principaux concepts.

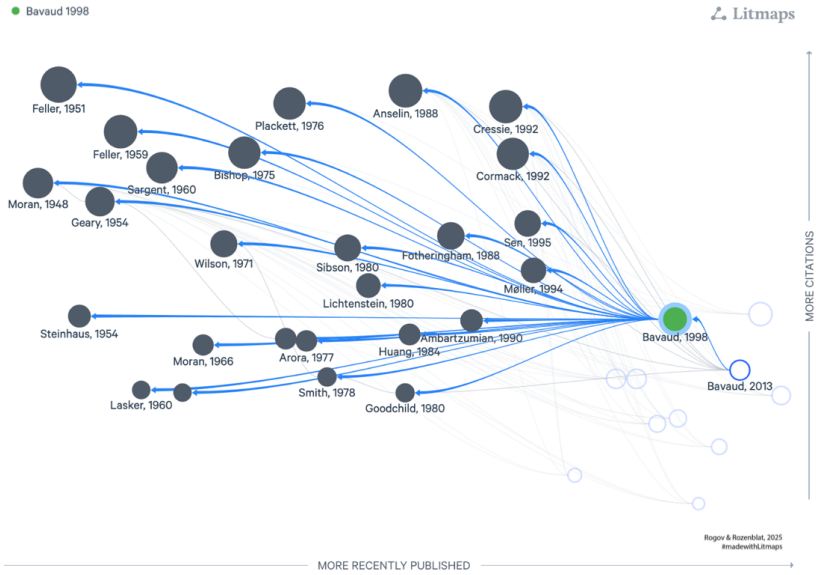


FIGURE 4 – Références bibliographiques de l'article Bavaud (1998).

Les 30 références bibliographiques sont cette fois fort différentes de celles de l'article de 1991 (Fig. 4). On y retrouve l'analyse spatiale avec Moran (1948, 1966) et Geary (1954), qui demeurent jusqu'à aujourd'hui parmi les auteurs majeurs de l'autocorrélation spatiale, mais aussi Wilson (1971) grand spécialiste des modèles gravitaires et de l'entropie en géographie. Puis des auteurs plus jeunes et aujourd'hui très reconnus dans la géographie théorique et quantitative comme Fotheringham & O'Kelly (1989), Goodchild & Smith (1980), Griffith & Anselin (1989),

Hordijk & Nijkamp (1977), Schwab & Smith (1985). Toutefois il y associe d'autres sources venant par exemple de la physique pour la géométrie intégrale (Ambartzumian, 1990; Santalo, 1976), ou des statistiques sociales pour l'équivalence des tableaux de contingence (Bishop et al., 1975; Caussinus, 1965).

Les termes et associations de ces références sont ici bien plus simples (Fig. 5). Trois principaux groupes émergent. La **Spatial Analysis** regroupe des termes utilisés en analyse spatiale (*spatial interaction*, *flows*, *theory*, *models*, *gravity model*, *spatial autocorrelation*) ; le groupe **Theories** concerne des réflexions plus théoriques (*analysis*, *applications*, *references*, *problem*, *alternative approaches*) ; Le groupe **Statistical physics** vient de deux textes en physique statistique (Bishop et al., 1975; Haberman, 1976). A noter que ces deux textes se renvoient l'un l'autre, puisque Haberman (1976) est la revue critique dans le journal « The Annals of Statistics » du livre de référence précédemment cité de Bishop et al. (1975). Un quatrième groupe de termes, **Statistics**, concerne les applications sur les données (*data*, *spatial data*, *statistics*, *index*) avec des références se rapportant plutôt à l'analyse spatiale comme Geary (1954). Arora & Brown (1977) ou Hordijk & Nijkamp (1977).

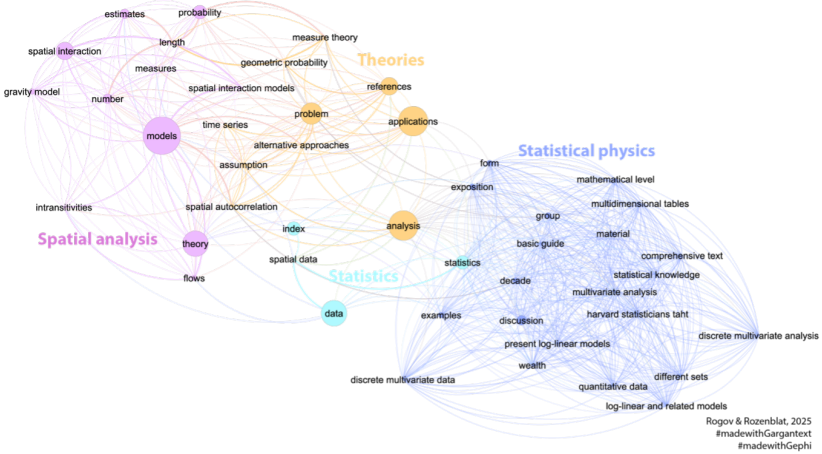


FIGURE 5 – Thématiques des références de l'article Bavaud (1998).

3.3 Célébrer la tradition dans une perspective d'innovation

Le troisième article que nous analysons est :

> Bavaud, F. (2011). On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28:297–314.

Cet article de 2011 apparaît à la suite de plusieurs années de publications sur les propriétés spatiales (concentration géographique, flux) (Bavaud, 2008a,b) ou les propriétés textuelles (Bavaud & Xanthos, 2005; Bavaud et al., 2006). Bien entendu, les recherches méthodologiques restent très présentes dans la décennie 2000, portant essentiellement sur des comparaisons d'efficacité des méthodes et algorithmes comme l'analyse factorielle des correspondances, le « power kernel » associé aux Support Vector Machines (SVM) (Bavaud et al., 2006) ou sur l'établissement formel des classifications floues (Bavaud, 2004).

Entre 2010 et 2011, François Bavaud revient plusieurs fois sur les transformations de Schoenberg (1938). Dans cet article de 2011 en particulier, il se plait à faire redécouvrir leurs propriétés qu'il compare aux kernels gaussiens utilisés en apprentissage automatique. Ainsi, évaluant l'efficacité de cette méthode traditionnelle de transformations de Schoenberg, basée sur les distances euclidiennes, à définir des covariances robustes, il démontre que cette approche est équivalente à celle des noyaux (kernels) et « sans doute plus intuitive » (p. 312).

Les références de cet article remontent pour six d'entre elles entre 1926 et 1941, puis six également entre 1960 et 1979, 22 entre 1980 et 1999, et 22 également de 2000 à 2010 (Fig. 6).

Les références de la première période 1926-1941 posent les bases conceptuelles historiques telles que : les fonctions monotones (Bernstein, 1929), l'intégration dans l'espace de Hilbert et les espaces métriques et fonctions définies positives (Schoenberg, 1938), les distances mutuelles (Young & Householder, 1938), les intégrales de Fourier (Neumann & Schoenberg, 1941).

La deuxième période 1960-1979 se réfère au *scaling* (Torgerson, 1961), aux propriétés de distance des racines latentes des méthodes vectorielles utilisées dans l'analyse multivariée (Gower, 1966) ou aux

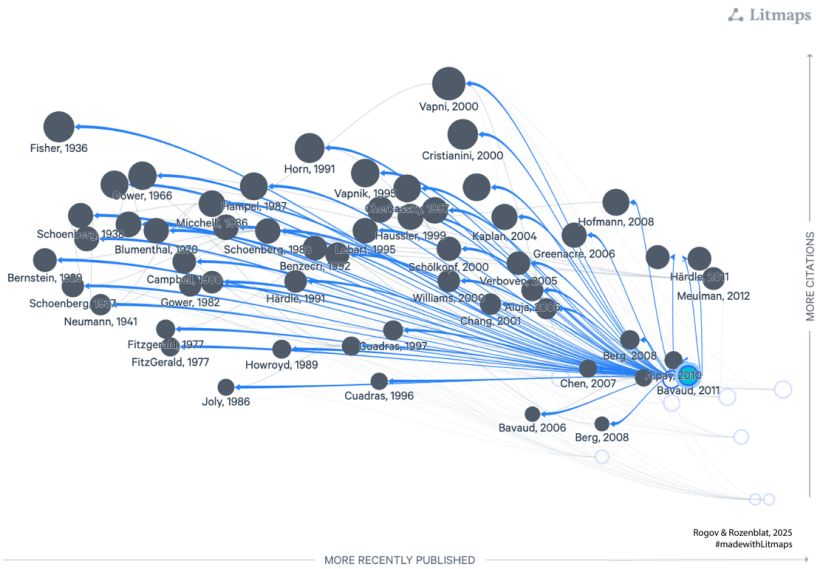


FIGURE 6 – Références bibliographiques de l'article Bavaud (2011).

puissances fractionnaires de Hadamard des matrices définies positives (Fitzgerald & Horn, 1977).

Les années 1980 et 1990 apportent les variogrammes (Christakos, 1984), des avancées sur les analyses des correspondances (Benzecri, 1992) et les puissances fractionnaires de Hadamard des matrices définies positives (Howroyd et al., 1989). Les années 1990 enrichissent également le scaling pondéré et multivarié (Borg & Groenen, 1996; Cuadras & Fortiana, 1996), l'interpolation des données spatiales et le « krigeage » (Stein, 1999) ou les noyaux de convolution (Haussler, 1999).

Quant aux années 2000, elles consolident les méthodes kernel dans les apprentissages automatiques (Hofmann et al., 2008) et apportent des sophistications dans les méthodes multivariées et de clustering (Bavaud, 2009).

Les termes lexicaux utilisés par toutes ces références soulignent six champs différents qui ne suivent pas d'ordre chronologique, mais plutôt

montrent des travaux anciens et récents qu'il combine dans chacun des domaines (Fig. 7).

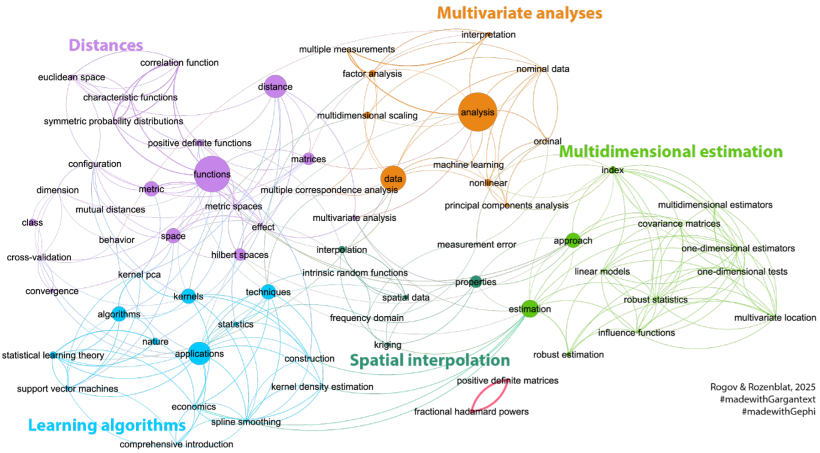


FIGURE 7 – Thématiques des références de l'article Bavaud (2011).

Les **Multivariate analyses** incluent l'analyse des correspondances multiples sur des données nominales ou ordinales et l'analyse en composantes principales et elles sont associées à l'apprentissage automatique (machine learning). La classe **Multidimensional estimations** associe les tests uni-dimensionnels ou multi-dimensionnels aux matrices de covariance. La classe **Distances** concerne les fonctions de distance et métriques spatiales tandis que la classe **Spatial interpolation** regroupe les méthodes d'interpolation et de kriging. La classe **Learning algorithms** identifie les applications d'algorithmes de noyaux (kernels) et les théories d'apprentissage statistiques.

Au total, l'article de 2011 de François Bavaud démontre l'ampleur de sa base théorique et méthodologique qui construit une certaine cohérence d'un panel de références fondatrices de différents domaines autant que de leurs avancées plus récentes. Il confirme la profondeur de sa réflexion qui se nourrit continuellement de chaque domaine en parallèle pour produire son cocktail original qui les combine.

4 Des champs scientifiques inspirés par les travaux de François Bavaud

Ainsi, l'originalité des travaux de François Bavaud tient dans la subtile combinaison de multiples champs scientifiques et nous pouvons faire l'hypothèse qu'à leur tour ils inspirent des champs scientifiques eux-mêmes variés. L'exploration des travaux citant les trois articles que nous avons précédemment étudiés va partiellement nous éclairer sur l'ampleur des champs scientifiques touchés et la nature de leurs impacts.

4.1 Une pierre dans l'édifice des modèles thermodynamiques

L'article de 1991 est cité 92 fois de manière très régulière jusqu'à aujourd'hui (Fig. 8).

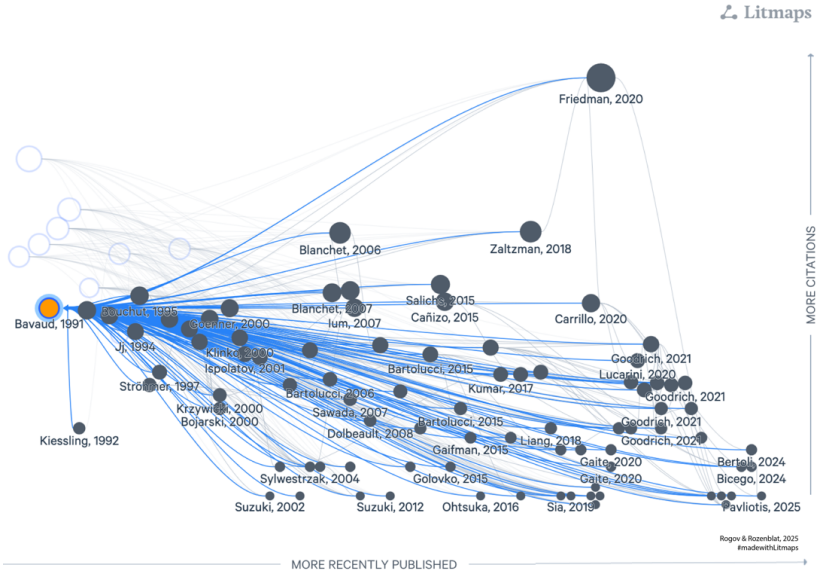


FIGURE 8 – Travaux citant l'article de Bavaud (1991).

Le domaine de la physique caractérise l'ensemble de ces articles (Fig.9) et en particulier la **Thermodynamics** qui concerne 31 articles dont ceux parmi les plus récents autour des transitions de phase, des énergies des systèmes dynamiques de particules, etc. (i.e. Bertoli et al.,

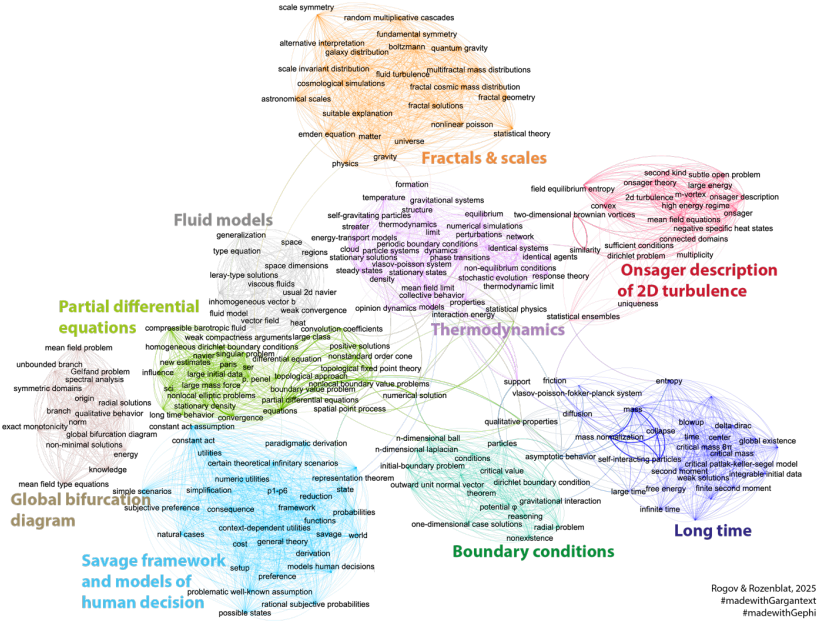


FIGURE 9 – Thématiques des travaux citant Bavaud (1991).

2025; Lucarini et al., 2020). Les termes de la classe des **Partial differential equations** qualifient 19 articles venant de champs différents. Certains de ces articles sont clairement ancrés dans cette classe, notamment ceux du champ des mathématiques comme Goodrich (2021). La plupart des citations viennent toutefois de la physique, comme Kiessling (1992) qui démontre une transition de phase du premier ordre dans les états de Gibbs pour une large classe d'interactions, ou Blanchet et al. (2006) qui s'attachent à prouver une inégalité entre l'énergie libre et sa dérivée temporelle dans le système de Keller-Segel (qui décrit le mouvement collectif de cellules attirées par une substance chimique et capables de l'émettre). Certains de ces articles contribuent également à former la classe des **Fluid models**.

D'autres articles appliquent les **Partial differential equations** dans des disciplines variées comme l'écologie (Hastings & Gross, 2012) ou la médecine sur la croissance des cellules cancéreuses

(Cui & Friedman, 2003). On trouve également des citations dans des domaines se focalisant sur des modèles spécifiques comme les **Onsager description of 2D turbulence** (Bartolucci & De Marchis, 2015; Bartolucci et al., 2018; Suzuki, 2012), **Savage framework and models of human decision** (Gaifman & Liu, 2015) qui associent également des systèmes auto-gravitants (Aly, 1994) que l'on retrouvera toutefois plus nombreux dans la classe **Boundary conditions** (par ex. Biler & Nadzieja, 1993; Soler et al., 1997). Un autre groupe de termes concerne les **Fractals & scales** en cosmologie notamment (Gaite, 2020; Miller et al., 2023). La classe **Long time** contient des termes insistant sur la dimension temporelle des modèles d'équation interagissant fortement avec les autres classes de termes (Blanchet et al., 2006, 2008; Bouchut & Dolbeault, 1995; Dolbeault, 1999; Toshpulatov, 2023). La classe **Global bifurcation diagram** ne partage aucun terme avec les autres classes bien que l'unique article formant cette classe (Bartolucci & Jevnikar, 2021) soit écrit par des co-auteurs que l'on retrouve dans les autres classes avec d'autres articles. Dans cet article de 2021, ils analysent le « comportement qualitatif du diagramme de bifurcation global de la branche non bornée des solutions du problème de Gelfand traversant l'origine (Gelfand, 1963, p. 1) » (traduction des auteurs) tout en utilisant la dynamique des particules en interaction avec elles-mêmes, qu'ils ont puisée dans l'article de François Bavaud de 1991. Car nul n'est jamais assez fort pour ce calcul.

4.2 Un pilier pour les approches spatio-temporelles des interactions

En se consacrant à l'approfondissement des propriétés des matrices pondérées, l'article de François Bavaud de 1998 est cité 229 fois et s'ouvre cette fois à des disciplines bien plus variées (Fig. 10). L'article est toujours cité par quelques physiciens, mais qui deviennent très minoritaires au profit des statisticiens, des géographes, des études de transport, des sciences régionales, des économistes, ou encore des sociologues.

Un grand point commun de toutes ces disciplines est la question des simulations spatiales à partir de données spatio-temporelles d'interaction avec la classe centrale **Spatial simulations** (Fig. 11). Cette

classe est formée par 57 articles insistant sur les *weighted matrices*, avec 10 textes traitant des *relationships*, 4 des *spatial interdependencies*, 3 des *spillovers*, et 13 des *distributions*. On trouve aussi des termes courants très fréquents dans les modélisations tels que *effets* (concerne 14 documents) et *impact* (16 documents).

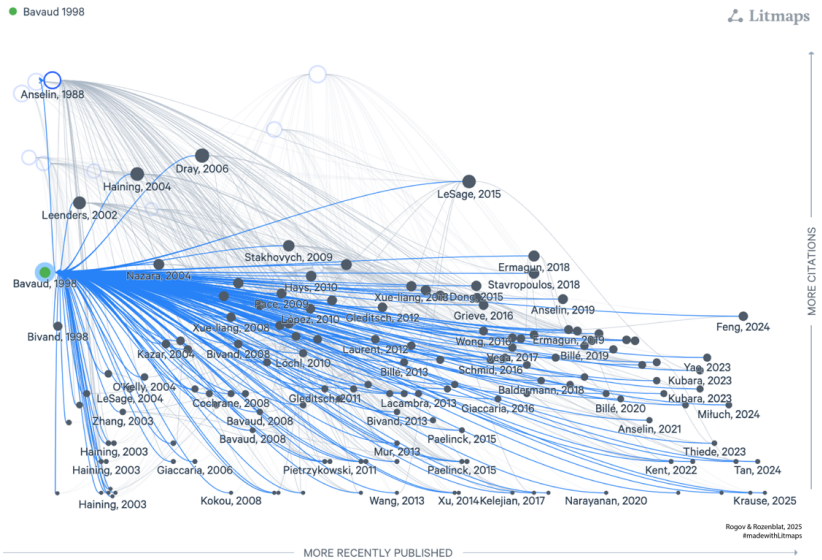


FIGURE 10 – Travaux citant l'article de Bavaud (1998).

Autour de ce groupe central « général », on trouve des thématiques plus spécialisées. Les articles contribuant au groupe des **Spatial autoregressive models** sont essentiellement publiés dans la revue *Geographical Analysis* (Dong & Harris, 2015; Ermagun & Levinson, 2018), *Papers in Regional Science* (Stakhovych & Bijmolt, 2009; Zhou & Lai, 2011) ou d'autres revues ou ouvrages de modélisation économique (Anselin, 2021; Krause & Kripfganz, 2025; Leenders, 2002; Scrucca, 2005). Un autre groupe de **Spatial econometric models**, n'est constitué que d'un article (Billé & Arbia, 2019) dédié à l'économie de la santé qui fait une revue des modèles à variables spatiales.

Le groupe **Networks & weighted matrices** rassemble des articles méthodologiques sur l'utilisation des mesures de centralité et des modèles

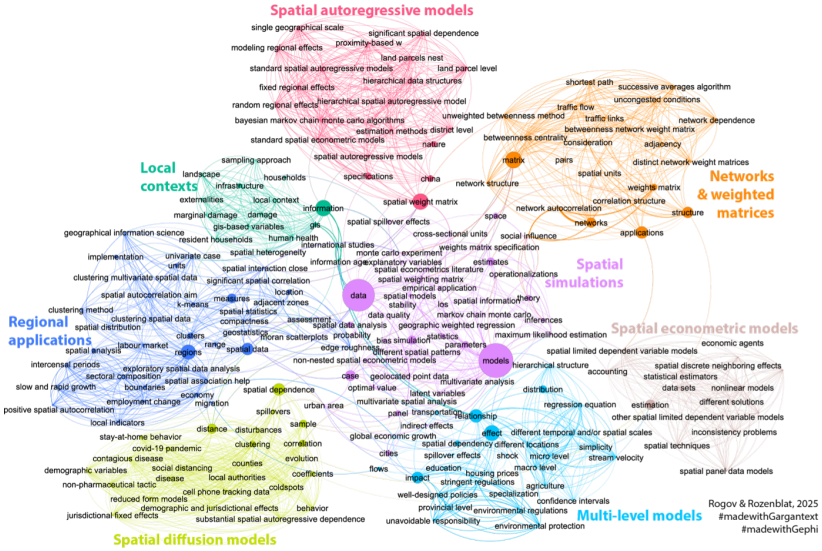


FIGURE 11 – Thématiques des travaux citant l'article de Bavaud (1998).

bayésiens sur les réseaux ou d'autocorrélation en physique, en économie ou en géographie (Dray, 2011; Ermagun & Levinson, 2018; Pace & LeSage, 2009; Stakhovych, 2010; Zhang, 2003; Zhang & Murayama, 2003). Contribuent également à cette classe des articles d'applications en économie (Billé & Catania, 2018), en géographie (Ermagun & Levinson, 2019; Zhu & Diao, 2020) ou en économie politique (Herrera Gómez et al., 2011). Le groupe **Spatial autoregressive models** réunit des termes des articles précédemment cités, mais particulièrement focalisés sur les algorithmes de modélisation et les méthodes d'estimation. Le groupe **Multi-level models** est constitué de deux articles d'écologie écrits par le même couple d'auteurs (Berk & de Leeuw, 2006; de Leeuw & Berk, 2003).

Les groupes **Regional applications** et **Spatial diffusion models** concernent davantage des statistiques appliquées dans les sciences sociales. Le premier groupe **Regional applications** réunit des clusterings dans les domaines du *marché du travail* (Baxendine et al., 2005; Scrucca, 2005), des relations entre la *migration* et la *convergence* internationale

(Østbye & Westerlund, 2007), du *shift & share* sur l'emploi (Cochrane & Poot, 2008), des études de proximité en écologie (Nelson & Robertson, 2012) ou propose une application *R* pour la modélisation des *données aréales* (Bivand et al., 2013). Le groupe **Spatial diffusion models** est formé par deux articles des mêmes auteurs sur les *déterminants démographiques, juridictionnels ou spatiaux* dans la *distance sociale* durant l'épidémie du *COVID-19* (Narayanan et al., 2020). Le groupe **Local contexts** concerne deux articles d'autres auteurs qui estiment les *effets décroissants de la distance* dans des contextes socio-économiques variables sur la *perception* des habitants de certaines *infrastructures énergétiques* (Giaccaria et al., 2010, 2016).

Au total, on remarque ici une grande diversité d'approches qui s'inspirent de cet article de François Bavaud de 1998 pour en tirer différents aspects : les graphes pondérés, les autocorrélations spatiales, les clustering multivariés de variables spatiales, les modèles de distance, la modélisation dynamique et multi-échelle. Cela révèle la richesse multidimensionnelle de cet article et explique son grand nombre de citations.

4.3 Un pont entre les anciennes théories et les data sciences

L'article de François Bavaud de 2011 est cité 24 fois (Fig. 12). Bien que moins cité que les deux autres articles, cet article opère un retour synthétique aux fondamentaux de la physique théorique et à leurs liens qu'il explicite dans les approches plus contemporaines. Les thématiques des articles qui le citent sont clairement articulées et les citations se réfèrent souvent à ces combinaisons de thématiques (Fig. 13).

A l'image de cet article de François Bavaud, plusieurs des documents le citant opèrent des ponts entre les différents champs conceptuels de la physique, des statistiques et de l'informatique, notamment perceptibles dans les termes généraux situés au milieu du graphe qui concernent un mélange entre **Computation and Mathematics** auquel se réfèrent des manuels (Yao, 2019), des ouvrages sur le risk management (Dionne & Koumou, 2018) ou sur l'étude de la perception de dieu par les enfants (Cocco & Ceré, 2023). Tous ces documents ont en commun d'articuler les différents champs et de les relier aux champs actuels des data sciences.

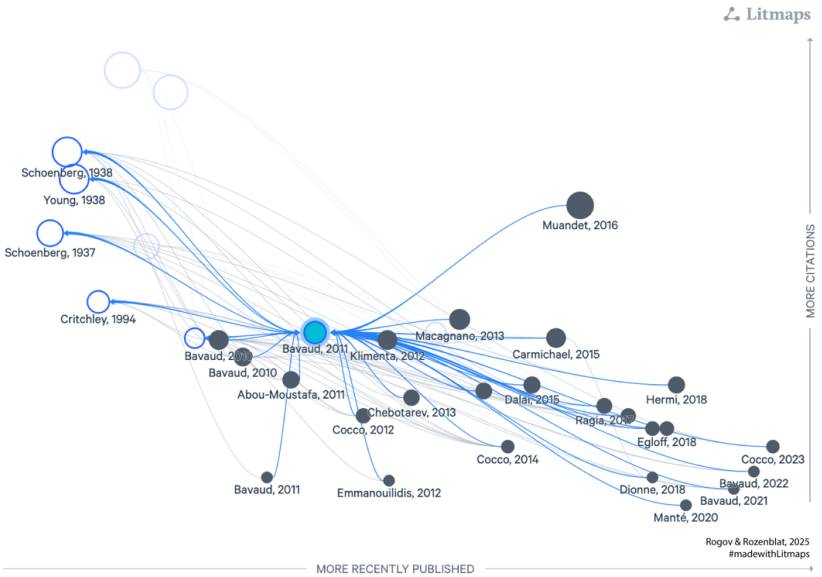


FIGURE 12 – Travaux citant l'article de Bavaud (2011).

L'article est cité spécifiquement sur les **Euclidian distances and clustering** se référant directement aux *Schoenberg transformations* (Alfakih, 2018), ou au *spectral clustering* (Abou-Moustafa et al., 2011). Les termes *similarities*, *entropy* et *multidimensional scaling* se retrouvent associés (Cocco, 2014; Macagnano & De Abreu, 2013) pour former le groupe **Similarities and entropy**. Des auteurs comme Muandet et al. (2017), revisitant les *kernels* dans le cadre du *machine learning*, contribuent au voisinage intime entre les groupes **Kernel and metrics of space** et **Unsupervised learning**.

Au total, cet article de « retour aux sources » semble boucler une boucle non définitive en reliant d'anciennes théories dans le cadre des approches contemporaines d'apprentissage : il s'agit de mettre à l'épreuve autant les théories précédentes que les nouvelles en confrontant leurs propriétés fondamentales. De ce point de vue, cet article est une étape importante constituant un pont robuste entre des paradigmes reposant fondamentalement tous sur des lois de probabilité

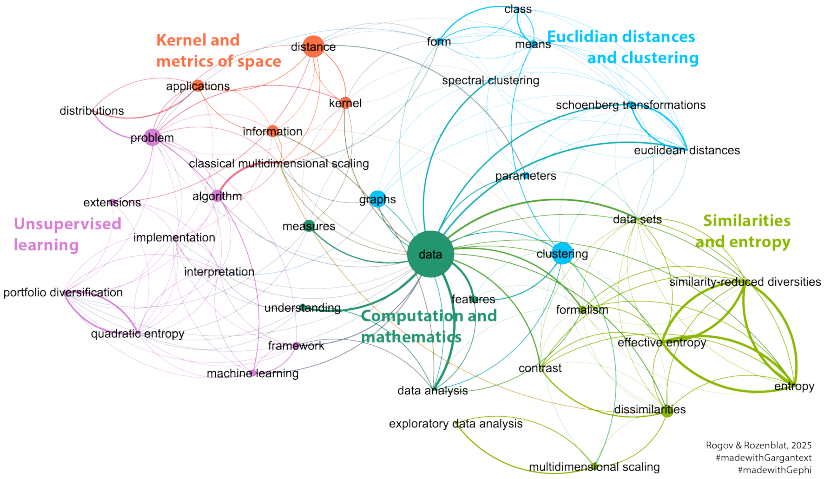


FIGURE 13 – Thématiques des travaux citant l'article de Bavaud (2011).

mais développant des algorithmes variés qui appellent à la réflexion que François Bavaud nous offre.

5 Conclusion

Ce chapitre, s'il est loin de représenter l'ensemble de la carrière de François Bavaud, a tenté d'en apporter quelques éclairages. On nous pardonnera notre incapacité à entrer dans toutes les subtilités des modélisations évoquées au cours de cette balade en « Bavaudie ». D'une modeste position de géographes quantitativistes « utilisateurs » de modèles, nous n'avons pu que souligner à quel point les bases théoriques en physique statistiques que François Bavaud offre à notre discipline, sont étendues bien au-delà de nos propres compétences, et combien les travaux qu'il a inspirés le sont encore bien davantage.

Toutefois notre démarche ne semble pas vaine : à partir de l'analyse bibliographique que nous avons menée, l'œuvre de François Bavaud apparait comme une élaboration de multiples ponts :

— Des ponts temporels entre des théories de *statistical mechanics* (Bavaud & Xanthos, 2002, p. 4 ; Bavaud, 2009, p. 74) qui l'ont précédé, qui l'ont accompagné et qui se développent encore : les

trois articles pris comme exemples sont encore abondamment cités aujourd'hui et le seront certainement encore dans le futur.

— Des ponts entre ces théories et les approches des distances et des dissimilarités à partir de graphes valués telles que développées en géographie, en analyse textuelle mais également en économie, en écologie, en science politique ou en sociologie.

Au-delà des ponts, les mélanges opérés par François Bavaud constituent une recherche intrinsèque des formules tant mathématiques qu'épistémologiques et verbales, à l'image de son sous-titre dans l'article de 1991, de « textes réchauffés et refroidis » (Bavaud, 2009, p. 74 ; Bavaud & Xanthos, 2002, p. 4). Derrière ce type de titres, on aperçoit un jeu libre des formules, des mots et des concepts en perpétuelle redéfinition et questionnements profonds sur les mécanismes sous-jacents de la pensée et du classement : penser avant de classer ou classer avant de penser (Perec, 2003) ? C'est un peu comme dans le *Traité du zen et de l'entretien des motocyclette*s (Pirsig, 1974) où l'auteur erre à la recherche d'une certaine beauté comprise tant dans le sens que dans la forme. Cette recherche active sans fin de la beauté des formules que François Bavaud nous donne à voir représente plus qu'une carrière : c'est un parcours de vie profitant pleinement de la « liberté académique » qui lui est donnée, que nous avons eu l'extrême chance de côtoyer et de partager, et pour cela nous l'en remercions profondément.

Références

Abou-Moustafa, K., Shah, M., De La Torre, F., & Ferrie, F. (2011). Relaxed exponential kernels for unsupervised learning. In *Joint Pattern Recognition Symposium*, pages 184–195, Berlin, Heidelberg. Springer.

Alfakih, A. Y. (2018). *Euclidean distance matrices and their applications in rigidity theory*. Springer, Cham.

Aly, J. J. (1994). Thermodynamics of a two-dimensional self-gravitating system. *Physical Review E*, 49(5):3771.

Ambartzumian, R. V. (1990). *Factorization calculus and geometric probability*. Cambridge University Press, 1ère édition.

Anselin, L. (2021). *Spatial models in econometric research*. Oxford University Press, Oxford.

Arora, S. S. & Brown, M. (1977). Alternative approaches to spatial autocorrelation: an improvement over current practice. *International Regional Science Review*, 2(1):67–78.

Bartolucci, D. & De Marchis, F. (2015). Supercritical mean field equations on convex domains and the Onsager's statistical description of two-dimensional turbulence. *Archive for Rational Mechanics and Analysis*, 217:525–570.

Bartolucci, D. & Jevnikar, A. (2021). On the global bifurcation diagram of the gelfand problem. *Analysis & PDE*, 14(8):2409–2426.

Bartolucci, D., Jevnikar, A., Lee, Y., & Yang, W. (2018). Non-degeneracy, mean field equations and the Onsager theory of 2D turbulence. *Archive for Rational Mechanics and Analysis*, 230:397–426.

Baumgartner, B. (1976). Thermodynamic limit of correlation functions in a system of gravitating fermions. *Communications in Mathematical Physics*, 48(3).

Bavaud, F. (1987). Statistical mechanics of viscoelasticity. *Journal of Statistical Physics*, 46(3/4):753–775.

Bavaud, F. (1989). Statistical mechanics of convex bodies. *Journal of Statistical Physics*, 57:1059–1068.

Bavaud, F. (1991). Equilibrium properties of the Vlasov functional: the generalized Poisson-Boltzmann-Emden equation. *Reviews of Modern Physics*, 63(1):129–149.

Bavaud, F. (1998). Models for spatial weights: a systematic look. *Geographical Analysis*, 30(2):153–171.

Bavaud, F. (2004). On the comparison and representation of fuzzy partitions. [Notes de cours non publiées]. Université de Lausanne.

Bavaud, F. (2008a). The endogenous analysis of flows, with applications to migrations, social mobility and opinion shifts. *Journal of Mathematical Sociology*, 32:239–266.

Bavaud, F. (2008b). Local concentrations. *Papers in Regional Science*, 87(3):357–371.

Bavaud, F. (2009). Aggregation invariance in general clustering approaches. *Advances in Data Analysis and Classification*, 3:205–225.

Bavaud, F. (2010a). Euclidean distances, soft and spectral clustering on weighted graphs. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 103–118. Springer Berlin Heidelberg.

Bavaud, F. (2010b). Les transformations de Schoenberg : propriétés et applications en analyse des données. In *Journées de statistique*, volume 42, Marseille.

Bavaud, F. (2011). On the schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28:297–314.

Bavaud, F. (2022). Similarity-reduced diversities: The effective entropy and the reduced entropy. *Journal of Classification*, 39(1):100–121.

Bavaud, F., Choquard, P., & Fontaine, J. R. (1986). Statistical mechanics of elastic moduli. *Journal of Statistical Physics*, 42:621–646.

Bavaud, F., Picca, D., & Curdy, B. (2006). Non-linear correspondence analysis in text retrieval: a kernel view. In *Proceedings of JADT, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*, pages 741–747.

Bavaud, F. & Xanthos, A. (2002). Thermodynamique et statistique textuelle: concepts et illustrations. In *Proceedings of JADT, 6èmes journées internationales d'analyse statistique des données textuelles*, pages 101–111.

Bavaud, F. & Xanthos, A. (2005). Markov associativities. *Journal of Quantitative Linguistics*, 12(2-3):123–137.

Baxendine, S., Cochrane, W., & Poot, J. (2005). Description and spatial analysis of employment change in new zealand regions 1986-2001. Technical report, University of Waikato, Population Studies Centre.

Benzecri (1992). *Correspondence analysis handbook*. Marcel Dekker, Inc., New York, NY.

Berk, R. A. & de Leeuw, J. (2006). Multilevel statistical models and ecological scaling. In Wu, J., Jones, K. B., Li, H., & Loucks, O. L. (éd.), *Scaling and Uncertainty Analysis in Ecology*, pages 67–88. Springer Netherlands, Dordrecht.

Bernstein, S. (1929). Sur les fonctions absolument monotones. *Acta Mathematica*, 52(1):1–66.

Bertoli, B., Goddard, B. D., & Pavliotis, G. A. (2025). Stability of stationary states for mean field models with multichromatic interaction potentials. *IMA Journal of Applied Mathematics*, page hxaf001.

Biler, P. & Nadzieja, T. (1993). Existence and nonexistence of solutions for a model of gravitational interaction of particles, i. *Colloquium Mathematicae*, 66(2):319–334.

Billé, A. G. & Arbia, G. (2019). Spatial limited dependent variable models: A review focused on specification, estimation, and health economics applications. *Journal of Economic Surveys*, 33(5):1531–1554.

Billé, A. G. & Catania, L. (2018). Dynamic spatial autoregressive models with time-varying spatial weighting matrices. Technical Report BEMPS55, Faculty of Economics and Management at the Free University of Bozen.

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W., Light, R. J., & Mosteller, F. (1975). *Discrete multivariate analysis: Theory and practice*. The MIT Press, Cambridge, MA.

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Modelling areal data*, pages 263–318. Springer, New York, NY.

Blanchet, A., Carrillo, J. A., & Masmoudi, N. (2008). Infinite time aggregation for the critical Patlak-Keller-Segel model in R^2. *Communications on Pure and Applied Mathematics*, 61(10):1449–1481.

Blanchet, A., Dolbeault, J., & Perthame, B. (2006). Two-dimensional Keller-Segel model: Optimal critical mass and qualitative properties of the solutions. *Electronic Journal of Differential Equations (EJDE)*, (44):33.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P10008.

Borg, I. & Groenen, P. J. F. (1996). *Modern multidimensional scaling*. Springer, New York, NY.

Bouchut, F. & Dolbeault, J. (1995). On long time asymptotics of the Vlasov-Fokker-Planck equation and of the Vlasov-Poisson-Fokker-Planck system with Coulombic and Newtonian potentials. *Differential and Integral Equations*, 8:587–514.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, 29:77–182.

Christakos, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research*, 20(2):251–265.

Cocco, C. (2014). *Typologies textuelles et partitions musicales: dissimilarités, classification et autocorrélation*. Thèse de doctorat, Université de Lausanne.

Cocco, C. & Ceré, R. (2023). Computer vision and mathematical methods used to analyse children's drawings of God(s). In *When Children Draw Gods: A Multicultural and Interdisciplinary Approach to Children's Representations of Supernatural Agents*, pages 213–244. Springer International Publishing, Cham.

Cochrane, W. & Poot, J. (2008). Forces of change: A dynamic shift-share and spatial analysis of employment change in New Zealand labour markets areas. *Studies in Regional Science*, 38(1):51–78.

Conlon, J. G. (1984). The ground state energy of a classical gas. *Communications in Mathematical Physics*, 94:439–458.

Cuadras, C. M. & Fortiana, J. (1996). Weighted continuous metric scaling. In Gupta, A. K. & Girko, V. L. (éd.), *Multidimensional Statistical Analysis and Theory of Random Matrices. Proceedings of the Sixth Eugene Lukacs Symposium, Bowling Green, Ohio, USA, 29–30 March 1996*, pages 27–40. De Gruyter, Berlin, Boston.

Cui, S. & Friedman, A. (2003). A free boundary problem for a singular system of differential equations: An application to a model of tumor growth. *Transactions of the American Mathematical Society*, 355(9):3537–3590.

Delanoë, A., Chavalarias, D., & Lobbé, Q. (2023). Gargantext: collaborative and decentralized libreware, version 0.0.7. Source Code.

Dionne, G. & Koumou, G. (2018). Machine learning and risk management: Svdd meets rqe. Technical Report 18-6.

Dolbeault, J. (1999). Free energy and solutions of the Vlasov-Poisson-Fokker-Planck system: external potential and confinement (large time behavior and steady states). *Journal de mathématiques pures et appliquées*, 78(2):121–157.

Dong, G. & Harris, R. (2015). Spatial autoregressive models for geographically hierarchical data structures. *Geographical Analysis*, 47(2):173–191.

Dray, S. (2011). A new perspective about Moran's coefficient: spatial autocorrelation as a linear regression problem. *Geographical Analysis*, 43(2):127–141.

Egloff, M. & Bavaud, F. (2018). Taking into account semantic similarities in correspondence analysis. In *Proceedings of the Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018)*, volume 2314 of *CEUR Workshop Proceedings*, pages 45–51.

Ermagun, A. & Levinson, D. (2018). An introduction to the network weight matrix. *Geographical Analysis*, 50(1):76–96.

Ermagun, A. & Levinson, D. M. (2019). Development and application of the network weight matrix to predict traffic flow for congested and uncongested conditions. *Environment and Planning B: Urban Analytics and City Science*, 46(9):1684–1705.

Fitzgerald, C. H. & Horn, R. A. (1977). On fractional Hadamard powers of positive definite matrices. *Journal of Mathematical Analysis and Applications*, 61(3):633–642.

Fotheringham, A. S. & O'Kelly, M. E. (1989). *Spatial interaction models: formulations and applications*, volume 1. Kluwer Academic Publishers, Dordrecht.

Gaifman, H. & Liu, Y. (2015). Context-dependent utilities: A solution to the problem of constant acts in Savage. In *Logic, Rationality, and Interaction: 5th International Workshop, LORI 2015, Taipei, Taiwan, October 28-30, 2015. Proceedings 5*, pages 90–101. Springer Berlin Heidelberg.

Gaite, J. (2020). Scale symmetry in the universe. *Symmetry*, 12(4):597.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146.

Gelfand, I. M. (1963). Some problems in the theory of quasi-linear equations. *American Mathematical Society Translations*, 29(2):295–381.

Giaccaria, S., Frontuto, V., & Dalmazzone, S. (2010). Who's afraid of power lines? merging survey and gis data to account for spatial heterogeneity. Technical Report 2, Department of Economics, "S. Cognetti de Martiis", Universita di Torino.

Giaccaria, S., Frontuto, V., & Dalmazzone, S. (2016). Valuing externalities from energy infrastructures through stated preferences: a geographically stratified sampling approach. *Applied Economics*, 48(56):5497–5512.

Goodchild, M. F. & Smith, T. R. (1980). Intransitivity the spatial interaction model and US migration streams. *Environment and Planning A*, 12(10).

Goodrich, C. S. (2021). A topological approach to nonlocal elliptic partial differential equations on an annulus. *Mathematische Nachrichten*, 294(2):286–309.

Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Griffith, D. A. & Anselin, L. (1989). Spatial econometrics: methods and models. *Economic Geography*, 65(2):160.

Haberman, S. J. (1976). Review of the book : Discrete multivariate analysis: Theory and practice. *The Annals of Statistics*, pages 817–820.

Hastings, A. & Gross, L. J. (2012). *Encyclopedia of theoretical ecology*. Number 4. University of California Press, Berkeley.

Haussler, D. (1999). Convolution kernels on discrete structures. http://ci.nii.ac.jp/naid/10015408231/.

Havlin, S. & Ben-Avraham, D. (1987). Diffusion in disordered media. *Advances in Physics*, 36(6):695–798.

Herrera Gómez, M., Mur Lacambra, J., & Ruiz Marín, M. (2011). Which spatial weighting matrix? An approach for model selection. *Asociación Argentina de Economía Política, Mar del Plata*.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220.

Hordijk, L. & Nijkamp, P. (1977). Dynamic models of spatial autocorrelation. *Environment and Planning A*, 9(5):505–519.

Howroyd, T. D., Upton, C. J. F., & Wood, W. W. (1989). Fractional Hadamard powers of positive definite matrices. *Real Analysis Exchange*, 15(1):21–25.

Kennedy, T. (1984). Mean field theory for Coulomb systems. *Journal of Statistical Physics*, 37:529–559.

Kiessling, M. K. (1990). A complementary thermodynamic limit for classical Coulomb matter. *Journal of Statistical Physics*, 59(5-6):1157–1186.

Kiessling, M. K. (1992). Finite-volume statistical mechanics of two-component Coulomb-like systems and the principle of macroscopic equivalence. *Communications in Mathematical Physics*, 146:311–331.

Kirkwood, J. G. & Monroe, E. (1941). Statistical mechanics of fusion. *The Journal of Chemical Physics*, 9(7):514–526.

Krause, M. & Kripfganz, S. (2025). Regional dependencies and local spillovers: Insights from commuter flows. *Journal of Regional Science*.

Leenders, R. T. A. (2002). The specification of weight structures in network autocorrelation models of social influence. Technical Report 02B09, University of Groningen, Research Institute SOM (Systems, Organisations and Management).

de Leeuw, J. & Berk, R. (2003). Final report to the EPA on multilevel models for generalization. Technical report.

Lucarini, V., Pavliotis, G. A., & Zagli, N. (2020). Response theory and phase transitions for the thermodynamic limit of interacting identical systems. *Proceedings of the Royal Society A*, 476(2244):20200688.

Macagnano, D. & De Abreu, G. T. F. (2013). Algebraic approach for robust localization with heterogeneous information. *IEEE Transactions on Wireless Communications*, 12(10):5334–5345.

Miller, B., Manfredi, G., Pirjol, D., & Rouet, J. L. (2023). From chaos to cosmology: Insights gained from 1D gravity. *Classical and Quantum Gravity*, 40(7):073001.

Moran, P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Moran, P. A. P. (1966). Measuring the length of a curve. *Biometrika*, 53(3-4):359–364.

Muandet, K., Fukumizu, K., Sriperumbudur, B., & Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Narayanan, R. P., Nordlund, J., Pace, R. K., & Ratnadiwakara, D. (2020). Demographic, jurisdictional, and spatial effects on social distancing in the United States during the COVID-19 pandemic. *Plos One*, 15(9):e0239572.

Neumann, J. & Schoenberg, I. J. (1941). Fourier integrals and metric geometry. *Transactions of the American Mathematical Society*, 50.

Pace, R. K. & LeSage, J. P. (2009). A sampling approach to estimate the log determinant used in spatial likelihood problems. *Journal of Geographical Systems*, 11(3):209–225.

Perec, G. (1991). *Cantatrix sopranica L. et autres écrits scientifiques*. La Librairie du XX^{e} siècle. Seuil, Paris.

Perec, G. (2003). *Penser/classer*. La librairie du XXIe siècle. Éd. du Seuil, Paris, nouvelle édition.

Pirsig, R. M. (1974). *Traité du zen et de l'entretien des motocyclettes*. The Bodley Head, London, réimpr. édition. Titre original: Zen and the art of motorcycle Maintenance: An inquiry into avlues, éd. française: Points.

Raveché, H. J. & Kayser, R. F. (1978). Towards a molecular theory of freezing: The equation of state and free energy from the first BBGKY equation. *The Journal of Chemical Physics*, 68(8):3632–3643.

Raveché, H. J. & Stuart, C. A. (1976). Bifurcation of solutions with crystalline symmetry. *Journal of Mathematical Physics*, 17(11):1949–1953.

Santalo, L. A. (1976). *Integral geometry and geometric probability*. Addison-Wesley, London.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44:522–536.

Schwab, M. G. & Smith, T. R. (1985). Functional invariance under spatial aggregation from continuous spatial interaction models. *Geographical Analysis*, 17(3):217–230.

Scrucca, L. (2005). Clustering multivariate spatial data based on local measures of spatial autocorrelation. *Quaderni del Dipartimento di Economia, Finanza e Statistica*, 20(1):11.

Soler, J., Carrillo, J. A., & Bonilla, L. L. (1997). Asymptotic behavior of an initial-boundary value problem for the Vlasov-Poisson-Fokker-Planck system. *SIAM Journal on Applied Mathematics*, 57(5):1343–1372.

Stakhovych, S. (2010). *Advances in spatial dependence modeling of consumer attitudes with Bayesian factor models*. Thèse de doctorat, University of Groningen.

Stakhovych, S. & Bijmolt, T. H. (2009). Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science*, 88(2):389–409.

Stein, M. (1999). *Interpolation of spatial data*. Springer.

Suzuki, T. (2012). Drift-diffusion model and 2D Brownian point vortices (modern approach and developments to Onsager's theory on statistical vortices). *Lectures at the Institute of Mathematical Analysis*, 1798:213–229.

Torgerson, W. S. (1961). Theory and methods of scaling. *Journal of the American Statistical Association*, 56(294):430–433.

Toshpulatov, G. (2023). Well-posedness and trend to equilibrium for the Vlasov-Poisson-Fokker-Planck system with a confining potential. *arXiv preprint arXiv:2310.12258*.

Wilson, A. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning A*, 3(1):1–32.

Yao, Y. (2019). A mathematical introduction to data science. [Notes de cours non publiées]. Hong Kong University of Science and Technology.

Young, G. & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22.

Zhang, C. (2003). Evaluation of irregular zone prominence based on geometric attributes. *Theory and Applications of GIS*, 11(1):53–59.

Zhang, C. & Murayama, Y. (2003). Evaluation on the prominences of irregular areas based on spatial weight matrices. *Geographical Review of Japan*, 76(11):777–787.

Zhou, O. & Lai, A. (2011). Suprasegmental contribution to the yelling reaction. experiments with stimulation and destruction. *Ztschr. f. d. ges. Neur. u. Psychiat.*, 56(130):631–677.

Zhu, Y. & Diao, M. (2020). Understanding the spatiotemporal patterns of public bicycle usage: A case study of Hangzhou, China. *International Journal of Sustainable Transportation*, 14(3):163–176.

Østbye, S. & Westerlund, O. (2007). Is migration important for regional convergence? comparative evidence for Norwegian and Swedish counties, 1980–2000. *Regional Studies*, 41(7):901–915.

Tabula Gratulatoria

En témoignage d'amitié et de reconnaissance

Adrian Spillmann
Agnieszka Soltysik Monnet
Alain Corbellari
Alberto Roncaccia
Allison Daley
Andrei Dobritsyn
Anita Auer
Anne Bielman Sánchez
Antoine Viredaz
Augustin Maillefer
Barbara Wahlen
Boris Vejdovsky
Brigitte Maire
Christelle Cocco
Christian Arnsperger
Christian Kaiser
Christina Stauffer
Christophe Lambiel
Danielle Chaperon
Danielle van Mal-Maeder
Davide Picca
Denis Renevey
Devis Tuia
Ekaterina Velmezova
Emmanuel Reynard
François Bussy
Frederic Herman
Frédéric Ratle
Gilles Merminod
Gilles Philippe

Grégoire Mariéthoz
Gretchen Walters
Guillaume Stern
Hans-Georg von Arburg
Irene Weber Henking
Jean-Luc Epard
Jérôme Meizoz
Joël Zufferey
Johanna Marin Carbonne
Karine Rossier
Kerria Grize
Kirsten Stirling
Ludivine Stofer
Magali Pétermann-Glaus
Manon Rosset
Maria Elena Castiello
Marianne Kilani-Schoch
Marie-Christelle Pierlot
Marie-Hélène Côté
Marj Tonini
Martin Müller
Maryam Kordi-Kaiser
Max Henking
Michel E. Fuchs
Michel Jaboyedoff
Michiel de Vaan
Nadia Spang Bovey
Neige Rochant
Niklas Linde
Oriane Martin
Pascale Deneulin
Patrick Rérat
Raphaël Bubloz
Raphaël Ceré
Remi Jolivet
Rémy Freymond
Rudolf Mahrer
Sebastian Pabst
Séverine Morel
Simone Albonico
Stefan Markus Schmalholz
Stéphanie Grand
Stéphanie Pichot
Stuart Lane
Théophile Emmanouilidis
Torsten Vennemann

Les Cahiers du CLSL
Centre de linguistique et des sciences du langage
CHF 20.- par numéro
En libre accès sur : www.cahiers-clsl.ch

Faculté des Lettres
Bâtiment Anthropole
Université de Lausanne
1015 – Lausanne, Suisse

ISBN 978-2-940607-19-8